Unsupervised Word Segmentation and Wordhood Assessment

Pierre Magistry

Université Paris Diderot (Paris 7)
École doctorale 132 : Sciences du Langage
U.F.R. de Linguistique

Numéro attribué par la bibliothèque

Thèse

Nouveau régime

Pour obtenir le grade de
Docteur en Sciences du Langage
Discipline: Linguistique Générale

Présentée et soutenue publiquement
par

Pierre MAGISTRY

Le 19 décembre 2013

Unsupervised Word Segmentation and Wordhood Assessment

The case for Mandarin Chinese

PhD supervisors:

Sylvain KAHANE,
Benoît SAGOT and Marie-Claude PARIS

Thesis Defence Committee:

PR. Sylvain Kahane (directeur)	Université Paris 10 Nanterre, MoDyCo
PR. Marie-Claude Paris (co-directrice)	Université Paris Diderot, LLF
DR. Benoit Sagot (co-directeur)	INRIA, ALPAGE
PR. 謝舒凱 Hsieh Shukai (pré-rapporteur)	National Taiwan University
PR. Yves Lepage (pré-rapporteur)	Waseda University
DR. Pierre Zweigenbaum (président)	CNRS, LIMSI

À Babcia

Remerciements

Je tiens tout d'abord à remercier mes encadrants, Sylvain Kahane, Marie-Claude Paris et Benoît Sagot qui m'ont fait confiance et m'ont accordé beaucoup de liberté dans mon travail tout en se montrant toujours très disponibles et d'excellents conseils. Depuis les années de licence et master où je suivais leur cours à Paris-Diderot et jusqu'aux derniers moments de la rédaction de ce mémoire, leurs influences sur les idées et les réalisations présentées ici sont majeures.

Je suis également très reconnaissant à Yves Lepage, Hsieh Shu-Kai et Pierre Zweigenbaum pour avoir accepté d'évaluer mon travail et d'être membres du jury. Je sais d'ores et déjà que les remarques qu'ils m'ont adressées seront d'une grande importance pour mes travaux à venir. Il me faut aussi remercier Rachel Bawden pour sa relecture attentive de mon anglais et ses commentaires détaillés qui m'ont permis d'améliorer la qualité du texte (M'étant fixé des délais un peu trop serrés, je n'ai pas pu lui fournir la totalité du texte et profiter pleinement de son aide. Il reste donc sûrement de trop nombreuses erreurs et maladresses qui sont entièrement de mon fait).

Avant d'être le produit du travail de ces dernières années, ce qui est présenté dans ce mémoire est le fruit de nombreuses rencontres avec des personnes m'ayant guidé ou accompagné et auxquelles je suis redevable. C'est d'abord Fabienne Marc qui m'a transmis de sa passion et de ses connaissances sur la langue et l'écriture chinoise. C'est elle aussi qui m'a suggéré de m'intéresser au traitement automatique des langues et m'a donné la motivation pour quitter mon travail de l'époque et reprendre les études à temps plein pour viser le doctorat.

Les années de master puis de doctorat au sein de l'équipe ALPAGE furent aussi instructives qu'agréables. La compagnie des autres doctorants, Charlotte, Chloé,

Corentin, Emmanuel, Enrique, Juliette, Luc, Marianne, Marion, Rosa et Valérie aura rendu ces années considérablement plus joyeuses et détendues que ne peut l'être la conduite d'une thèse dans d'autres circonstances. Ça va me manquer.
Au delà des murs de Paris 7, les fidèles des «Informels» ont aussi contribué à ma bonne humeur et dans le même temps à élargir mes horizons linguistiques à d'autres langues et d'autres questionnements. Merci surtout à Nicolas, Olivier, Ilaine, Ivan, Ji-hye. J'espère que ces petites réunions tout comme nos amitiés dureront encore longtemps !

Mon travail doit aussi beaucoup aux deux années passées à Taïwan entre le master et le doctorat. J'y ai poursuivi mes études de linguistique à L'Université Nationale de Taïwan et eu mes premières expériences de publication sous la direction de Cheung Hintat qui m'a initié aux recherches en acquisition du langage.

Ce séjour à Taïwan fut l'occasion de s'ouvrir à la réalité linguistique du terrain sinophone dans sa variation et sa complexité. C'est surtout Yoann qui m'a judicieusement poussé dans ce sens. Cette ouverture amène de nouvelles questions trop peu présentes dans ce mémoire, mais ça viendra vite ! Et son obsession à chercher partout la clef des champs m'aura au moins donner envie de creuser un peu le mien.

De façon plus surprenante, c'est aussi à Taïwan que j'ai eu l'occasion de rencontrer, de travailler et de me lier d'amitié avec les Toulousains de la «Proxteam». Ils m'ont ainsi montré que nous vivons bel et bien dans un petit monde ! Merci à Ben, Bruno, Karine, Manu, Yann et Yannick qui tout en travaillant à d'autres projets m'ont souvent offert un regard exterieur et une oreille attentive sur mes travaux. Ils m'ont permis par là de les clarifier. J'espère très vite voir les premières applications pratiques de mon travail en l'intégrant aux leurs !(oui je sais, ça ne tient plus qu'à moi…)

Enfin, je voudrais remercier mes parents qui m'ont regardé poursuivre ces projets un peu fous de thèse et d'allers et venues à l'autre bout du monde. Avec un tel comportement un fils s'attirerait à coup sûr les foudres de Confucius, mais eux ne m'en ont jamais tenu rigueur. Ils m'y ont même encouragé et m'ont soutenu au fil de toutes ces années !

Contents

List of Tables

List of Figures

Foreword

After decades of research in Natural Language Processing and a hundred years of works aiming to describe Modern Standard Chinese, it may seem surprising, or unnecessary to come back to the question of *wordhood*. However, up until present there appears to be no fully satisfying definition of the minimal units for linguistic description and for language processing that also fits the needs of quantitative linguistics.

In my own previous work, the inadequacy of existing segmentation systems to my needs for linguistic studies has been a major issue. During my work on the quantitative analysis of the productivity of morphological rules (Magistry, 2008), state-of-the art segmentation systems were not an option because they performed (and still perform) poorly when dealing with the items I was targeting: rare occurrences of long words. I had to restrict my experiments to a manually segmented corpus and this was the main limitation of my work.

During my stay in Taiwan, I had the great opportunity to participate in the ongoing Franco-Taiwanese project called M3 "Model and Measurement of Meaning" (Desalle et al., 2010; Magistry et al., 2009, among others). This project compared results from psycho-linguistic experiments on language acquisition to computational linguistic models of the lexicon. The lexical models were based on synonymy graphs of Mandarin and French. Questions related to the segmentation of Mandarin took a large part of the discussions in the first joint meeting in which I participated where both the Taiwanese and French teams were present.

The issue revolved around the choice of vertices for the Mandarin lexical graph. The question "what is a word in Chinese?" has been around for decades, and it was inevitably asked once again. It felt like this question could neither be avoided nor definitively answered. Building the Mandarin graph was one of the tasks of the M3 project. The French graph on the other hand was already available before the

beginning of the project and was not a subject of discussion in itself. It had been compiled by merging multiple synonymy dictionaries, therefore it inherited from traditional French dictionary entries the decisions on what constitutes a vertex in the graph. As a consequence the French graph included many so-called Multi-Word Expressions (MWE). But discussing the relevance of the presence of such expressions in the graph was outside the scope of the project.

It appeared to me that this striking difference in term of how lively the debates about *wordhood* were on each side of the table was not so closely related to typological differences between French and Mandarin. On the contrary, it was mostly the result of differences in terms of lexicographic traditions and trust in the available resources.

In view of what has been said above, I felt I had to address this question in my dissertation.

General linguistics offers many definitions and theories concerning its minimal units. But in practice, applying these theories to large-scale corpora often ends up in endless debate and disagreement among human experts. The current practice in Chinese word segmentation is to provide guidelines with heuristics that allow expert annotators to reach a satisfying level of consistency in the manual segmentation of the training corpus. Then supervised machine learning is used to mechanise the segmentation process in order to segment unseen data. In doing so, we purposely leave behind linguistic felicity.

On the other hand, I believe it possible to refine the definitions and segmentation procedures inherited from early works in structural linguistics, in order to provide segmentation algorithms which would not need manual annotation but would extract the implicit structure from raw language data. This is precisely the purpose of unsupervised machine learning. The resulting segmentation will thus depend on the corpus being segmented and may vary from one corpus to another, but this seems to be a desirable property which corresponds to psycho-linguistic findings. More importantly, the results will be perfectly reproducible. This contrasts with supervised learning whose segmentation heavily depends on training data and whose reproducibility is limited by the reproducibility of the manual annotations.

If an unsupervised segmentation system is designed to follow sound principles from structural linguistics, its output can be expected to be equally sound. This is

what is explored throughout this dissertation.

As the ideas presented here are the result of an unquantifiable number of interactions over the last few years, when I was working in very inspiring and diverse environments with many people to whom I feel indebted, I have decided to write the remainder of this dissertation using the plural pronoun “we”, even though “we” is not a well defined and constant group of people, especially from one chapter to another.

Part I.

Background

Chapter 1

Introduction

This dissertation addresses the problem of the characterisation and recognition of linguistic minimal units, in particular minimal syntactic units. These are the minimal building blocks for syntax, and their internal structure is described by morphology. The present work thus falls right in the middle of these two levels of analysis.

We propose a linguistically sound yet practical method to measure the syntactic autonomy of the forms present in a corpus and propose an algorithm to perform the segmentation of utterances into relevant segmentation units using the proposed measure of autonomy.

1.1. The Need for Segmentation

Speech and writing[1] are essentially observable as a linear (unidimensional) sequence of phonemes or characters. However, to analyse utterances, virtually all formal theories of syntax require some kind of multidimensional representation to describe the relations between possibly non adjacent sub-sequences of characters.

For example, for the linear utterance *Mon vieil ami chante cette jolie chanson* 'My old friend sings this beautiful song', Tesnière (1959) gives the dependency analysis reported in Figure 1.1 (the *stemma* under Tesnière's terminology). In this trivial case, segmenting on the space characters does yield satisfying units. However if we translate this simple sentence into Chinese, we may produce the sequence of characters 我老朋友唱這首好聽的歌. To be able to propose the analysis illustrated in Figure 1.2, we need a segmentation that cannot be trivially derived from a graphical

[1] Sign languages offer a different challenge.

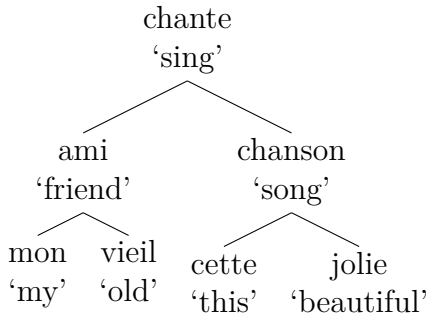


Figure 1.1.: **Stemma for French (from Tesnière, page 15)**

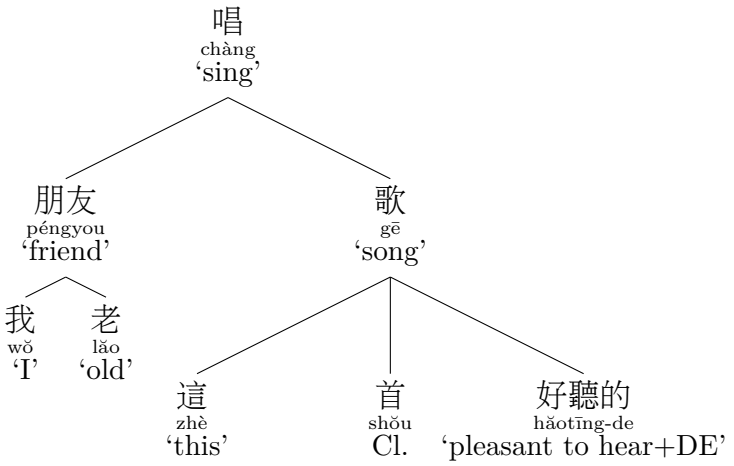


Figure 1.2.: **Stemma for Chinese**

clue. The segmentation into 我•老•朋友•唱•這•首•好聽的•歌[2] is a prerequisite to a syntactic analysis. The case for French may also not be as simple as it first seems because of the so-called “multiword expressions” discussed in chapter 4.

(Tesnière, 1959, chapter 10) makes interesting statements about the notion of the word (the vertices of his stemma). He considers that sentences are not to be defined starting from its words. Quite the opposite, it is the words that are to be defined from the sentence. He states that “one can not define the word in itself, but only from the *breaks* that delimit its beginning and its end” and describes these “breaks” as being:

- partial, they never completely break the speech;

[2]Throughout the dissertation, we indicate explicitly the segmentation using middle dots• and question the segmentation at a specific position using ∘

- variable, “the cuts they dig are not of an equal depth”;
- relative, they can only be measured with respect to one another.

Tesnière considers that these cuts are “not only unclear, but also impossible to clear up. Therefore the notion of the word is essentially elusive.”

We agree with Tesnière's description, but we believe that this notion can be made less elusive with the help of computational linguistics. The corpora now available and computation power allow us to estimate the “depth” of the cuts based on statistics drawn from unannotated corpora and thus to attempt an empirical, data-driven definition of the word.

We will see that a single notion of “word” is not sufficient to describe the variety of phenomena we can observe in terms of syntactic autonomy. Therefore we prefer to talk about ‘syntactic units’ or ‘segmentation units’.

We shall try to characterise these units (chapters 2 to 5), underline the limitations of current practices in NLP (chapters 6 and 7) and propose a new method to perform the segmentation of the linear input (part 2).

1.2. Language, Speech and Corpus

Saussure et al. (1916) makes the distinction between *la langue* ‘language’ and *la parole* ‘speech’, contrasting the “set of necessary conventions adopted by the social body so as to permit the usage of the faculty of language among individuals” with “the act of the individual putting his faculty into practice”.

The language data that we can observe and collect (*i.e.*, the corpus) is a collection of “speeches” that individuals have consciously crafted to pursue a specific communicative goal. They did so freely to a large extent but they nevertheless had to follow language conventions to build understandable messages.

We work under the hypothesis that by doing so, they let the language (*la langue*) impose its structure on their speech (*leur parole*) in a way that is statistically observable and that makes human speech distinguishable from a random sequence of symbols or sounds.

Unlike most previous works, we are not designing a segmentation system which would explicitly follow a specific linguistic theory. We rather take the position of the observer and try to see what the distributions we find in corpora can reveal about linguistic units and boundaries. We check afterwards whether it matches our theoretical position (presented in section 3.3).

Our task naturally points to lexical units but we proceed bottom-up: we select abstract principles that should define linguistically relevant boundaries, implement them in a segmentation system and observe the output to see which aspects of linguistic theories are captured.

1.3. Language Acquisition vs Language Description

We work under the hypothesis that the structure of the language can emerge from a wide range of speech varieties. We consider this assumption reasonable because the situation our system will be facing is to some extent comparable with the situation in which children acquiring their first language are placed.

However, this thesis is not an attempt to model the actual process of language acquisition (unlike related work presented in section 7.2.2). We do not claim any psycholinguistic felicity regarding the segmentation procedure. Nevertheless it is very likely that the clues useful to our system are somehow related to the clues used by infants. Describing this relationship is outside the scope of our study but would grant a cognitive plausibility to the output of our system.

Rather than language acquisition modelling, our work can be considered as a first step towards a fully mechanised structural linguistic analysis in which language structure is induced from raw data.

1.4. Computational Linguistics and Natural Language Processing

We evaluate our work following the current practice in the Natural Language Processing community. We do so because it is the only available way to compare our system to other systems and prove its output sensible. Nevertheless, we underline

the limitations of such an evaluation in chapter 6 and section 7.2.6. Our purpose is the computational and automatic study of language, we do not have any specific language technology application in mind (in this dissertation at least), apart from pursuing the automatic analysis and description of a language.

1.5. Working on Written Data

As our work is achieved based on written materials, it is worth noting from the start that provided a system to convert an acoustic signal into a sequence of symbols is available, everything that is presented in this dissertation should be applicable.

This being said, we do not subscribe to the idea that writings are only a secondary encoding of a language which is of lesser importance for linguistics studies compared to speech data. From Saussure's 1916's statement that 'language and writing are two distinct systems of signs" we draw the conclusion that writing is a system of its own that deserves to be studied too.

As recent works in corpus linguistics and NLP massively target written material, further discussion about the relevance of such study seems unnecessary. A more comprehensive discussion on this point can be found in (Harris, 2005). The exact relation between spoken and written languages is also outside the scope of our work.

1.6. The Choice of Mandarin Chinese as a Case Study

This dissertation focuses on the current standard form of Mandarin Chinese, or Modern Standard Chinese (referred to as MSC throughout this dissertation). Some aspects of what we are referring to with the name MSC are presented in chapter 2 with further references for details.

We did not choose to work on MSC because the ideas and algorithms proposed here only apply to this language. On the contrary, we would argue that this work can be extended to any language and script. But MSC exhibits interesting properties that make our point easier to make than it would have been on many other languages (French for example).

Firstly, MSC in its standard written form does not explicitly mark any kind of unit that resembles what we intuitively call a "word" in French or English. Without

any intuitive notion of "word", the difficulty to define a segmentation unit is more striking (especially to NLP practitioners). This is discussed in details in chapter 2. It results in a large body of previous works in linguistics (chapter 3), a variety of existing segmentation systems for MSC as well as well-established evaluation practices (chapters 6 and 7).

Secondly, MSC presents no inflection system. This greatly simplifies our problem. A proper treatment of inflection is left for a future work.

Thirdly, the use of Chinese Script for MSC provides an interesting graphical unit, larger than a Latin letter, that correspond to a syllable with a certain degree of semantic disambiguation: a given syllable can be transcribed by various distinct characters, depending on the intended meaning (chapter 2 and 3).

1.7. Outline of the Dissertation

This is dissertation is organised as follows:

The first part focuses on the linguistic aspects of our work and its relations to previous works in both linguistics and Natural Language Processing.

The next chapter provides an overview of what is traditionally considered as a *word* outside linguistics and the equivalent unit, the character (字 zì) for Chinese. It explains why a more subtle analysis of the minimal units is required in linguistics.

Chapter 3 discusses the linguistic definitions and criteria from the Chinese linguistics literature and describes the position we adopt in our work in section 3.3.

In chapter 4, we argue that the issues raised for Chinese are not limited to Modern Standard Chinese (MSC) but can be related to that raised by the presence of Multi-Word Expressions (MWE) in any language.

Chapter 5 explains the specificities of a study solely based on raw corpus data, compared to traditional linguistic studies that rely also on introspection and elicitation.

Chapter 6 provides a short history of the establishment of Chinese Word Segmentation as a typical task for NLP researches. It discusses the available corpora for training and evaluation and their guidelines.

Chapter 7 presents different approaches to the problem of word segmentation, with a specific focus on the unsupervised systems more closely related to our work.

The second part of this dissertation presents our novel approach to the assessment

of the combinatoric *autonomy* of forms, the first required step towards a fully unsupervised and automatic assessment of wordhood.

Chapter 8 presents our improvements of the reformulation of a Harrissian hypothesis concerning the segmentation for it to be operational on corpus data.

Chapter 9 stresses the relevance and the importance of preprocessing, which significantly increases the quality of our segmentation.

Chapter 10 reports experiments on refining our system by using the Minimum Description Length principle. It shows mixed results that on the one hand allow us to achieve state-of-the-art performance, but on the other hand limit the genericity and the clarity of our approach.

Chapter 11 presents a new method to evaluate the quality of a segmentation. It relies on dependency structures annotated in a Treebank and provides a finer-grain evaluation. In this chapter we also provide a visualisation technique and a way to formulate our results as a probabilistic model. This allows us to better understand the behaviour of our system and to confirm its relevance for our long-term goals.

Chapter 12 concludes this dissertation with preliminary results from auxiliary experiments on other languages. It also provides insights about our future research directions.

Chapter 2

A First Naive Approach: the Orthographic and Sociological Word

The notion of "*word*" may sound like an intuitive one. For English or French, it may seem to be part of the speaker's intuition that it is futile to question. We will see that this impression is misleading and that our intuitive notion of the word is mostly a cultural artefact based on orthography. Regardless of which language is analysed, a strict definition of minimal units for text processing is not trivial to find.

2.1. Orthographic Word and Writing Systems

To begin, we claim that what appears as intuitive to the speakers of languages such as French or English (written with the Latin alphabet) is the correspondence of two different possible definitions of *word*. Namely what Chao (1968) and Packard (2000) call the *sociological word* and the *orthographic word*.

The *sociological word* is the cultural item present in many social activities such as boardgames. It is the basic unit of the traditional dictionaries we are used to. The *orthographic word* is the unit defined by typography. In modern Latin script, they are typically surrounded by *spaces* or other punctuation marks. This, in fact, is part of the orthography and has to be learned explicitly. Consider examples (1) and (2) in French.

(1) a. ***au*** *milieu*
at-the middle
'in the middle'

b. ***au****tour*
at-the-periphery
'around'
c. *aumilieu

(2) a. *à seule* ***fin*** *de*
at only end of
'for the sole purpose of'
b. *a****fin*** *de*
at-end of
'in order to'
c. #à fin de

Many things could be said to justify the presence or the absence of white-spaces in such contrasting examples. However a justification will hardly convince us that those are not essentially conventions that are learnt at school. Correct spacing is thus part of the *orthography*, just like the use of any punctuation mark. The *orthographic word* arises from tradition and the explicit normalisation of writing practices. For practical reasons (it is technically trivial to split on space characters to turn a string of characters into a sequence of tokens), they are also the basic processing unit of most of the NLP applications and research for French and English.

Now if we look at the Chinese Script, the situation seems totally different. The set of punctuation marks available to write in MSC does not include a white-space or an equivalent way to systematically mark boundaries. On the other hand, each Chinese character is written in a (invisible) square of the same size and they are written with a constant distance between two characters. This makes each character easy to isolate from the others (even when handwritten in cursive style). It follows that Chinese characters are a natural candidate for the *orthographic word* in Chinese script. Interestingly, the Chinese character is also the *sociological word* when people write using Chinese script. The differences begin to appear when the Chinese characters are used as a basic unit for linguistic analysis and NLP.

When dealing with MSC, a more subtle treatment to select minimal processing units is required for linguistic analysis and for many NLP tasks. Interestingly, this has not always been well accepted. DeFrancis (1984) advocates a non strictly character-

based approach to study MSC, he explains in his chapter on "*the Monosyllabic Myth*" that the cultural status of the Chinese character was a source of confusion for early analysts. DeFrancis traces this confusion back to early missionary works that state that in Chinese, "words, syllable and written symbols are the same". He then explains the need to distinguish between free, semi-bound and completely bound characters: "These three categories are roughly comparable in English to the free form *teach*, the semi-bound form *er* in *teacher* and *preacher*, and the completely bound forms *cor* and *al* in *coral*. The first two categories are morphemes, the third is not, as is the case also with their counterparts in Chinese".

DeFrancis also provides some statistics based on a two hundred characters sample from the "Concise Dictionary of Spoken Chinese" Chao and Yang (1962), one of the few dictionaries that explicitly label characters as free or (semi)bound. Completely bound forms are noted as forming part of a unique fixed combination of characters. The figures he gives are reported in Table 2.1.

Free	44%
Semi-bound	45%
Completely Bound	11%

Table 2.1.: **Degree of Syllabic freedom, from DeFrancis (1984, p. 185)**

Using one of the manually segmented corpora that we shall present in more detail in chapter 6, we computed similar statistics. As we base our statistics on corpus, we are able to account for ambiguity (characters that have both free and semi-bound uses).

The training part of the corpus provided by the Peking University (hereafter the PKU corpus, which we will describe in details in chapter 6) contains 4574 distinct Chinese characters for a total of 1.6M occurrences. 2886 characters (63%) can occur as a free word and 4425 (97%) can occur inside a multi-character word (as a semi-bound or completely bound form). 2737 characters (60%) have at least one use as a free morpheme and one use as a bound morpheme. Only 515 characters (11%) appear only as free morphemes and 1688 as bounded morpheme only. 1654 are completely bound (they always appear in the same multi-character word). If we consider all the 1.6M occurrences of characters, only 350,000 (22%) are occurrences of free morphemes and 78% are bound morphemes inside a larger word.

The differences between our figures and those reported by DeFrancis can be explained by the difference in data source. DeFrancis used a "concise dictionary"

in which many low frequency words and characters present in our corpus were left out by Chao and Yang. The other factor is that the authors of the dictionary have probably considered that if a character has a usage as a free morpheme, it is not considered as possible semi-bound form. If we discard low frequency words and remove the free forms from the list of ambiguous forms, we obtain figures similar to DeFrancis's.

In any case, DeFrancis's position against a monosyllabic treatment of MSC still holds with our figures. The use of characters as the basic units of processing is equivalent to the assumption that each character is totally free. This is inadequate for the majority of the characters in a corpus.

For the case of French, assuming that every *orthographic word* is free may also be wrong in the case of so-called "Multi Word Expressions" (MWE). Chapter 4 will focus on MWE, and points out that examples like *au milieu de* could or should be considered as a complex but single unit (a preposition). We can also consider example (3), also a MWE in which *fur* is not free and can even be argued to have no meaning in synchrony (it is *Completely Bound* in the terms of DeFrancis, just like *cor* in *coral*). In synchrony, this word only appears in the idiomatic expression *au fur et à mesure*, it is therefore impossible to induce a meaning from its use without looking at an etymological dictionary. Most native speakers of French will know the expression *au fur et à mesure* but not know that *fur* used to mean *price* then *rate* (ending up as a near-synonym for *mesure*).

(3) *au* ***fur*** *et* *à* *mesure*
at-the ? and at measure
'gradually'

So it seems to us that the two situations are not fundamentally different. Starting from *orthographic words* in French is also error-prone. Only the proportion of erroneous forms is not the same. Constant et al. (2011) notes that in the French Treebank (Abeillé et al., 2003), 15% of the *orthographic words* are annotated as belonging to a larger MWE (and most of them may be free forms in other contexts). These 15% correspond to the 78% we found in our manually segmented Chinese corpus if we consider multi-character words as "Multi (orthographic) Word Expressions" in MSC.

The respective success and failure of the naive approach for NLP in French and

Chinese is due to the respective proportions of non-free *orthographic* forms. The need for a more subtle processing has thus been felt since the beginning of Chinese NLP whereas MWE are less crucial in NLP for languages written in Latin script and are ignored in many practical applications.

To have a complete picture, it is important to note that this proportion of non-free forms in *orthographic words* is not constant. We shall mention that although orthographic norms seem more prescriptive than descriptive in their relation to language, the need for word boundary punctuation marks was felt by writers long before any kind of standardisation. If writers feel the need for it, it must be part of their intuition about the language. It is worth taking a glimpse into the evolution of writing practices.

2.2. A Short History of Orthographic Word Boundaries in Writing

In NLP literature on Chinese Word Segmentation, it is common to read that "unlike English, Chinese has no word boundaries". A lot of confusion underlies this kind of statement. First of all, it seems important to distinguish the language from the script used to write it. A given script may be used to represent multiple languages in slightly different ways and a given language may be represented by multiple scripts. For example, the Latin script can be used to represent English, French, German, and so on. The Chinese script is used for modern Mandarin Chinese but also for classical Chinese, Japanese, Taiwanese and Cantonese amongst others. On the other hand, Chinese characters in Vietnamese, Korean or Taiwanese can be or have been replaced by other scripts.[1] Some languages including Japanese, Korean and Taiwanese are also written with Chinese script mixed with other scripts.

The relationship between a script and a language is subject to evolution through time, so are the scripts themselves. This evolution may affect the explicit marking of boundaries. It is worth mentioning that ancient Greek and Latin were initially written in *scriptio continua*, *i.e.*, without any word boundary markers. The habit of writing explicit boundaries came about progressively, first with a middle dot (a

[1] Interestingly, it has been so with different strategies. For example romanised Vietnamese kept the same *orthographic word* as Vietnamese in Chinese script but romanised Taiwanese join these units with a hyphen and mark larger *orthographic words* with spaces.

·) and later with spaces. This habit for Latin script became the common usage in Europe only during the second half the the first millennium (see Drillon (1991)).

Although DeFrancis claims that the equivalence between words, syllables and written symbols in Chinese has always be wrong, we can argue that it has been "less wrong" than it is nowadays. At some point, the situation for Chinese was comparable with the current situation for *orthographic words* in French (Latin script), that is to say, with an orthographic word boundary mark (the space between two Chinese characters) that corresponds well to a syntactic and semantic unit. At that time, Chinese script was used to represent Archaic and Classical Chinese. Compare examples (4) in Classical Chinese (a quotation from Confucius's Analects and (5), a possible traduction in MSC.

(4) 季路 問 事 鬼神 。 子 曰 : 「 未 能 事 人
Jilu ask serve ghost-spirits . master say : " not-yet can serve man
, 焉 能 事 鬼 ? 」
, how can serve ghosts ?"
'Ji Lu asked about serving the spirits of the dead. The Master said, "While you are not able to serve men, how can you serve their spirits?"'

(5) 季路 問 如何 事奉 鬼神 。 孔子 回答 説 : 「 不 能
Jìlù wén rúhé shìfèng guǐshén kǒngzǐ huídá shuō bù néng
JiLu ask how serve gods . Confucius answer saying :" not can
事奉 人 , 怎麼 能 事奉 鬼神 啊 ? 」
shìfèng rén zénme néng shìfèng guǐshěng a
serve man , how can serve gods PART ?"

We can spot some expressions involving multiple characters in the Classical Chinese example, but such expressions were not as pervasive as in its MSC counterpart.

With the evolution of Classical Chinese, more and more multi-characters expressions with non compositional meanings or with no syntactic freedom appeared. Those could mostly be considered as "multiword expressions". As the proportion of those expressions increased, the boundaries between linguistically relevant expressions became more and more fuzzy.

More recently, it is a political event that caused the end of explicit word boundary

marking in "Chinese" (as a script): the shift from vehicular to vernacular languages for official written materials and the establishment of Modern Standard Mandarin Chinese as the national language of the Republic of China in a reform movement that started in 1913. For more details about the historical aspects of that shift, on can refer to (Norman, 1988) and (Kaske, 2008). To concur with DeFrancis's *Monosyllabic Myth*, we can stress that the first event in that direction in 1913 was not explicitly about building a national language, but about unifying the reading of Chinese characters. It was called the 讀音統一會 dúyīn tǒngyī huì *Commission on the Unification of Pronunciation* (for more historical and sociolinguistics details, see Chen (1999)).

This is indicative that even during the shift from vehicular to vernacular as the main written language, the Chinese character remained the *sociological word*.

The only aspects in which the reforms during the 20th century affected the script itself are the graphic simplification of some characters (only in the People's Republic of China), and the introduction of certain punctuation marks (which are indeed word boundary markers,[2] but not as pervasive as spaces can be in other scripts and languages).

With regard to our subject, this sociolinguistic evolution is the root of the confusion between Chinese Script and Standard Mandarin and the frequent unclear statement that "Chinese has no word boundaries".

The national language defined later was mostly based on the vernacular languages spoken in the North of China which are as distant from classical Chinese as many modern Romance languages can be from Latin. We shall note here that a large variety of vernacular languages spoken in China (including Mandarin) already had various traditions of writing before the institutionalisation of Mandarin Chinese as the national language and its standardisation. They were written both using Chinese characters and romanised characters (under the influence of missionaries, see section 12.2.1 for Taiwanese Hokkien as an example). When written in Chinese characters, no clear orthographic words larger than one character were observable. Romanisation however, did provide such larger units. Written vernacular Mandarin using Chinese characters also served as the basis to define the lexicon of Modern Standard Mandarin Chinese. However, influences of the various local vernacular languages can still be found both in oral and written languages.

[2] With the exception of the semicomma 、 used to form coordinated lists and which can be used to coordinate characters at the sub-word level, examples can be found in section 3.2.1.

This historical background is crucial for any work on Chinese languages, but the full picture is outside the scope of the present research. We just want to stress that *word* boundaries were first more clearly marked in texts written with Chinese characters. But those marks, or strictly speaking the function of those marks, slowly disappeared with language evolution and a strong and sudden change in language policies concerning the official written language led to the present situation.

To conclude this short incursion into the history of writing systems, we follow Drillon in claiming that spacing characters are part of the punctuation and what is commonly referred to as a *word* is what Packard (2000) describes as the *orthographic word*. It matches with the *sociological word* in both Latin and Chinese scripts, but it doesn't match linguistically relevant units to the same extent in the two scripts. In fact, this orthographic and sociological unit seems to be more resistant to changes than the linguistically relevant units that are to be defined.

2.3. Towards a Linguistically Sound Analysis

This mismatch between the *orthographic word* and a larger morpho-syntactic unit in MSC led many scholar to deny the existence or the relevance of *words* in Chinese. The distinction between characters (字 zì) and a larger word-like unit (詞 cí) was made in 1907 by Zhang Shizhao, but the relevance of this level of analysis was still contested years later. One reason is often the confusion between Old Chinese and Modern Standard Chinese at a time when MSC was still under the process of standardisation. This is especially visible in dictionaries where the definitions at the character level do not typically distinguish meaning of a free form and etymology of a character. There is often a lack of distinction between information about the actual language in use in synchrony and the knowledge drawn from philology. Another reason to deny the relevance of a *word*-like level was the difficulty to design a precise set of tests that would yield consistent analysis. See for example the position of Chao Yuanren in his grammar (Chao (1968), as cited by Duanmu (1998)):

> "Not every language has a kind of unit which behaves in most (if not all) respects as does the unit called 'word' ... It is therefore a matter of fiat and not a question of fact whether to apply the word 'word' to a type of subunit in the Chinese sentence."

Nowadays, the relevance of a *word*-like morpho-syntactic unit is widely acknowledged, even though a fully satisfying and widely accepted definition for it is still lacking. The set of usable tests to isolate and classify wordforms in MSC has been well described in the literature (Huang, 1984; Duanmu, 1998; Packard, 2000; Nguyen, 2006). We review the criteria that are the most relevant for our discussion in the next chapter. In Chapter 5 we discuss the usability of those linguistic criteria in a corpus study with a high degree of automation. Next we will explain how these tests relate to general linguistics to sketch out how our work can be extended to other languages. This will stress the need for unsupervised segmentation and lead us to a discussion on Chinese Word Segmentation in NLP, in chapter 6.

Chapter 3

The Word in Chinese Linguistics

This chapter reviews a selection of previous works that provide the traditional methodology to delimit and qualify words in MSC.

Chinese script does not mark relevant boundaries in MSC. On top of that, MSC has very few inflections. Spaces or other equivalent punctuation marks and inflection paradigms are the two main indicators of wordhood and word classes (*parts of speech*) in the languages that use the Latin alphabet. For example, English verbs can be characterised by a bound root which combines with a closed set of inflectional affixes and English nouns will typically take a suffix *-s* to mark a plural form. These two important clues are not available to analyse MSC, hence the need for a specific set of criteria and tests to distinguish *words*, phrases, affixes and clitics.

Methodology and criteria can be found in (Kratochvíl, 1967; Huang, 1984; Duanmu, 1998; Packard, 2000; Nguyen, 2006). This chapter revolves mostly around Kratochvíl, Duanmu and N'Guyen's proposals.

We also describe our adaptation of the description of minimal units exemplified with French in (Kahane, 2008) to MSC (Magistry, 2008).

Duanmu and N'Guyen both provide a clear review of the variety of tests that have been proposed to distinguish between the the various types of units. Duanmu does so to demonstrate consistency between a selection of morpho-syntactic criteria and a phonological definition of wordhood. His morpho-syntax is based on constituent analysis. On the other hand, N'Guyen aims at defining the guidelines to write an "Explanatory Combinatorial Dictionary" (Mel'čuk and Polguère, 1987) in the dependency-based Meaning-Text Theory (MTT) (Mel'čuk, 1988, 1994). We contrast the two approaches to stress important issues for our own purpose.

Kratochvíl (1967) gave some guidelines towards a description of MSC *word classes*. It is worth mentioning that Kratochvíl only outlined the methodology and explicitly refused to provide a comprehensive description of those classes because of the socio-linguistic context at the time. He claims that such work may have easily been taken as a prescriptive normalisation of MSC even if such a pioneering work would have necessarily been flawed in one way or another.

He first suggests the need for a system of word classes because it allows for a more concise description of the language and he starts with the following example:

(1) 我們 也 去
wǒmen *yě* *qù*
We also go
'We go too'

from which he extracts three commutation paradigms

- X: { 我 *wǒ* 'I', 你 *nǐ* 'you', 王博士 *Wáng bóshì* 'doctor Wang', 剛才 *gāngcái* 來 *lái* 的 *de* 那 *nà* 個 *ge* 人 *rén* 'the person that just arrived', …}
- Y: { 當然 *dāngrán* 'naturally', 慢慢地 *mànmànde* 'slowly', 等 *děng* 一會兒 *yīhuì'r* 就 *jiù* 'just in a moment', …}
- Z: { 來 *lái* 'come', 高興 *gāoxìng* 'happy', 跟他 *gēn tā* 借了 *jièle* 不少 *bùshǎo* 的 *de* 錢 *qián* 'borrowed a lot of money from him', …}

in the context of example (1), the first item can commute with any member of class X, the second with any member of class Y and the third with any member of class Z. Note however that not all X Y Z combinations are possible.

Kratochvíl calls the contexts of possible commutation "*substitution frames*". For the three aforementioned classes, these are :

(2) a. *X* 也 去
yě *qù*
X also go

b. 我們 *Y* 去
wǒmen *qù*
we Y go

c. 我們 也 *Z*
wǒmen *yě*
we also Z

The members of the classes proposed by Kratochvíl are of various kinds, ranging from *grammatical affixes* (including clitics) to *phrases* (XP).

To find out what the minimal forms are, Kratochvíl suggests proceeding by *reduction*. *Reduction* is comparable to *strong autonomy* in the MTT framework proposed in Mel'čuk (1994) that N'Guyen follows. In this example, 去 qù 'go' can be used alone with the same semantic contribution, for example as an answer to the question "*will you go?*". This would not be true for the first character 我 wǒ 'I' which is the first person singular pronoun whereas 我們 wǒmen is plural, both are different answers to the question "*who will go?*". The second character 們 men is a bound form that cannot occur alone, therefore 我們 wǒmen 'we' is a minimal autonomous unit. For expressions that cannot occur alone such as adverbial 也 yě 'also', Kratochvíl argues that the fact that they can be minimally suppressed during the reduction process without affecting the contribution of the other element to the sentence like in (3) can justify their treatment as wordforms.[1]

(3) 我們 去
wǒmen *qù*
We go
'We too'

This is for a sketch of Kratochvíl's procedure, let us now turn to the various types of unit that can be the outcome of the analysis.

3.1. Words, Phrases, Affixes and Clitics

***Words* and Affixes** In Chinese linguistics literature, a *word* is usually defined as the atomic unit of syntax. Duanmu describes it as an expression that can occupy an X^0 position in the X-Bar theory.[2] In dependency grammars such as MTT, it is a vertex of the graphs describing dependencies in surface syntax.

Chinese words can be made up of a single character or composed of multiple characters. Internal characters are divided into roots and affixes. The distinction between roots and affixes is a matter of productivity. Affixes are or have been sufficiently productive to form morphological families, which are sets of wordforms that are all derived from the same affixation rule. Bound roots may be used in multiple words with related meanings but without the regularity that would enable

[1] In this example, Kratochvíl seems to consider that 們 affects the meaning of the pronoun more than 也 'also' affects the meaning of the verb. We believe that other insights, such as the diversity of items with wich they combine is needed to reach the same conclusion.

[2] In the X-bar theory, The X^0 position is occupied by a lexical head to which is associated a part-of-speech X. It "projects" into a phrase XP

us to define a lexical rule. As productivity in morphology is typically a graded phenomenon (Baayen, 1992; Magistry, 2008; Arcodia and Basciano, 2012), we expect the boundary between word formative affixes and bound root compounds to be fuzzy. However this does not question the wordhood of the resulting composed expression. Under Kratochvíl terms, affixes can be either *word formative affixes* or *grammatical affixes*. Neither can occur alone but he distinguishes them because the former combine with bound roots and the latter combine with free wordforms.

However, Kratochvíl's *grammatical affixes* include a variety of items. Some are affixes such as 了 (le) (an aspectual marker for which it can be argued that it is a flexion of a verb), but others are considered as clitics or prepositions by others, for example the marker of potentiality 得 (de) is analysed as a clitic by N'Guyen. Some others are even segmented as grammatical wordforms by most of the segmentation guidelines such as the negation marker 不 (bù). (Those three items are given as example by Kratochvíl).

In more recent linguistic works, Kratochvíl's *grammatical affixes* are subdivided into clitics, affixes and phrasal-affixes Zwicky and Pullum (1983); Miller (1992). In our work, we call "affixes" what Kratochvíl calls "word formative affixes" and subdivide his "grammatical affixes" into "clitics", "prepositions" and "phrasal affixes".

Phrases, clitics and phrasal affixes Phrases are the projection of a lexical head in generative frameworks, or a head and its dependants for dependency grammars. In any case, they are a combination of words, grammatical words and clitics. The rules involved in these combinations have to be highly productive.

We reserve the word "clitics" to denote bound forms lacking prosodic structure and attached to their phonological host postlexically (as the selection of the host depends on the syntactic structure at a phrase level). For MSC this definition of clitics limits its use to a closed set of morphemes whose tone can be neutralised. We will use the term "phrasal affixes" for other bound form whose position is constrained at a phrase level. We also distinguish phrasal affixes from *grammatical words* such as prepositions. The difference is on a syntactic level; both are non-autonomous but grammatical words are the head of a phrase, which is not the case for phrasal affixes.

3.2. A Review of Commonly Used Wordhood Criteria

Both Duanmu and N'Guyen follow a methodology that is similar to Kratochvíl's, but they give a more detailed account of the way they make a decision for the unclear cases. Such cases typically arise when full *reduction* is impossible. They describe a set of criteria that is widely used in the literature and select those they consider the most relevant.

3.2.1. Conjunction Reduction

Conjunction reduction is the process of factorising a part shared by two coordinated elements. Huang (1984) suggests that such a reduction can be done with coordinated phrases but not with coordinated words.
Example (4) shows an acceptable reduction which demonstrates the autonomy of 書 shū *book*. It contrasts with (5) where 車 chē *-vehicle* is a bound morpheme.

(4) a. 舊 的 書 跟 新 的 書
jiù de shū gēn xīn de shū
old DE book and new DE book
'old books and new books'

b. 舊 跟 新 的 書
jiù gēn xīn de shū
old and new DE book
'old and new books'

(5) a. 火車 跟 汽車
huǒchē gēn qìchē
fire-vehicle and gaz-vehicle
'train and automobile'

b. *火 跟 汽車
huǒ gēn qìchē
'fire and automobile'

Regarding this test, Duanmu follows the criticisms raised by Dai (1992) and limits its applicability to the detection of bound forms. As some compounds may exhibit regular syntactic pattern, the impossibility of performing a reduction points to a wordform but its possibility does not imply that the tested form is a phrase.[3].

[3]Some languages allow for the conjunction reduction of affixes as in German *In- und Ausland*. Similar examples can be found in MSC such as 國内外 guónèiwài *country-inside-outside* only without a conjunction :

N'Guyen uses this test in a similar manner.

3.2.2. Freedom of parts

The Freedom of parts criterion states that if a component of an expression is a bound form, then the expression is a wordform.

For Duanmu, just like conjunction reduction, this test only applies when identifying wordforms. Effectively, two free forms can be combined in a single wordform (a compound word), as exemplified by (6) where 雞 jī *chicken* is a free wordform and 鴨 yā *duck* is a bound morpheme that requires the noun-formative suffix 子 zi to be used as a noun meaning *duck*. It is inappropriate to reach divergent conclusions for the two expressions in this example. Therefore Freedom of parts is used by Duanmu to conclude that 鴨蛋 yādàn is a wordform, but it says nothing about 雞蛋 jīdàn .

(6) a. 雞 蛋
jī *dàn*
chicken egg

b. 鴨 蛋
yā *dàn*
duck egg

N'Guyen rejects this criterion as unusable because it is cyclic: one needs to know that 鴨 yā is a bound form to conclude that 鴨蛋 yādàn is a wordform. But if we decide to treat 鴨 蛋 yā dàn as a phrase, then 鴨 yā becomes a free wordform without any contradiction.

We can add that this criterion cannot be applied with phrasal affixes. The result of the combination of a phrase and a phrasal affix cannot be considered as a word. If applied with word-level affixes only, we agree with N'Guyen's criticism.

3.2.3. Semantic Composition

This test is also of limited use to Duanmu and is rejected by N'Guyen. According to Chao (1968), once we establish that the parts of an expression are free, we can check whether its meaning is compositional. Expressions with compositional meaning should be treated as phrases and expressions with non-compositional meaning as

國内 guónèi 跟 gēn 國外 guówài but * 國内 guónèi 跟 gēn 外 wài

wordforms.

However, when applied *as-is* this criterion leads to the conclusion that all Multi-word Expressions (MWE) should be regarded as words. This would lead to contradictions with other criteria, especially *adverbial modification*, *XP substitution* and *productivity*. Since many MWE allow for a certain level of variability and modification. We will give a more detailed description of various kinds of MWE in chapter 4.

Because of this, Duanmu regards this criterion as no more than an indication and N'Guyen rejects its use.

3.2.4. Exocentric Structure

Exocentric constructions are expressions with an apparent inner syntactic structure but whose distribution is not the one expected for such a construction. For example, *je ne sais quoi* (lit. 'I don't know what') is a well formed sentence in French (headed by a verb) but in some contexts, it can be used as a noun, in which case it is often written with hyphens *un je-ne-sais-quoi*. This nominal use of the expression was even borrowed as such into English.

An expression which has an exocentric structure is regarded as a compound by Duanmu. N'Guyen defines the test by adding the condition that the considered exocentric structure not be attested as productive in the language. He gives the following example in French

(7) *Je tue quiconque me contredit*
I kill whoever myself contradict
'I kill whoever contradicts me'

where *[quiconque me contredit]* is a nominal phrase headed by a verb. Such a construction with a non-referent relative pronoun is productive in French and should not be regarded as a compound.

If we follow the exocentric structure criterion without such a precision, this would lead to the conclusion that *[quiconque me contredit]* is a compound wordform, which would be unfortunate as it conflicts with other criteria (*productivity, XP substitution, insertion, adverbial modification*).

3.2.5. Adverbial Modification

Duanmu retains the Adverbial modification criterion from Fan (1958) who remarks that in *[A* 的 *N]* structures, A can typically take an adverb of degree whereas *[A N]* does not allow such an adverb to be present. He gives the following examples.

(8) a. 新 的 書
xīn de shū
new DE book
'a new book'

b. 很 新 的 書
hěn xīn de shū
very new DE book
'a very new book'

c. 更 新 的 書
gēng xīn de shū
more new DE book
'a newer book'

d. 最 新 的 書
zuì xīn de shū
most new DE book
'the newest book'

(9) a. 新 書
xīn shū
new book
'a new book'

b. *很 新 書
hěn xīn shū
very new book

c. *更 新 書
gēng xīn shū
more new book

d. *最 新 書
zuì xīn shū
most new book

Duanmu explains this contrast by suggesting that *[Adv Adj]* is always a phrase and as such cannot occur inside a compound. On the other hand, *[Adj N]* are compounds.

N'Guyen adopts a similar criterion which is more general.

> "S_1S_2 is a wordform if at least one of its constituents loses its ability to accept modifiers when it is combined with the other constituent. On the

other hand, it is a phrase if its constituents still accept modifiers."

Note that constituents of idiomatic expressions may or may not lose their ability to accept modifiers. This will be discussed in the next chapter.

3.2.6. XP Subsitution

If one considers all *[Adv Adj]* to be phrases, Adverbial modification can be seen as a specific case of *XP Substitution* where the Adj is replaced by a phrase. The original test also comes from Fan (1958) who notes that N in *[M* 的 *N]* can be substitued by a typical NP such as *[Numeral Classifier N]*. Example (10) illustrates this substitution.

(10) a. 新 的 書
xīn de shū
new DE book
'a new book'

b. 新 的 三 本 書
xīn de sān běn shū
new DE three Cl. book
'three books that are new'

c. 新 書
xīn shū
new book
'a new book'

d. *新 三 本 書
xīn sān běn shū
new three Cl. book

Conclusions are consistent with the Adverb modification criterion.

3.2.7. Productivity Criterion

Phrases are supposed to be derivable from the phrase construction rules of the grammar. Such rules are defined to be productive, that is to say to be able to create a large number of phrases, including new, unseen ones.

If an expression has an internal structure that does not correspond to a productive rule in the grammar, it will be treated as a wordform (by both Duanmu and N'Guyen). Previous *[Adj N]* wordforms are an example of such cases as $X \rightarrow Adj\ N$ cannot be considered as a productive syntactic rule for MSC (unlike English). An other example given by N'Guyen is

(11) 自從
zìcóng
since-from
'since'

which could be analysed as $P \rightarrow PP$. Such a rule would not be productive in Mandarin and 自從 zìcóng is therefore treated as a wordform.

Note that this criterion allows the definition of wordforms but it cannot prove that an expression is a phrase. A wordform may have an internal structure that corresponds to a productive rule.

3.2.8. Syllable Count

Lü (1979) proposes to rely on the length of the expression and of its components.

> "The word in the mind of the average speaker is a sound-meaning unit that is not too long and not too complicated, about the size of a word in the dictionary entry."

An earlier version by Lu et al. (1964) applies only to *[N N]* expressions and states that if at least one component has a length of one character, the expression should be regarded as a word. If on the other hand the expression is made up of two components of length 2, it should be considered a phrase.

Duanmu rejects this criterion by showing that in many cases it conflicts with the others. However many examples given by Duanmu could be considered as multi-word expressions rather than long compounds words. We will discuss them in the next chapter.

It is worth mentioning that this criterion is still a good heuristic that is explicitly used in segmentation guidelines of large corpora for NLP. We will give more details about those in chapter 4.

3.2.9. Insertion

The insertion test was proposed in very early works and is subject to various disagreements. It was proposed by Chao (1968) and rejected by Huang (1984) and Packard (2000). Both Duanmu and N'Guyen judge it useful but they use it in a

different ways. (Note however that neither of them include it in their final procedure.) This test consists of inserting an item between two parts of an expression. Insertion is possible providing that:

- The resulting expression has the same structure as the original.
- The resulting expression has the same meaning as the original.

Duanmu stresses the difficulty in deciding whether two structures or meanings are identical or not and he therefore chose not to use this test. This is particularly true for the insertion of 的 (de), which is often used as an example. He also objects that this test only applies in one way: if the insertion is impossible, then the expression is probably a word. Otherwise nothing can be inferred. In this statement, he adopts the treatment of Verbal Resultative with potential 得 (de) as compound verbs from Chao (1968), exemplified by (12) and argues that similar cases of two compounds with an insertion can be found in English nominal compounds like *evening class* and *evening chemistry class*[4].

(12) a. 我 看見 他
wǒ *kàn-jiàn* *tā*
I look-perceive him
'I see him'

b. 我 看得見 他
wǒ *kàn-de-jiàn* *tā*
I look-DE-perceive him
'I can see him'

N'Guyen disagrees with this analysis and treats Verb-Resultative Constructions as phrases. He argues that the first verb still accepts modifiers and that the insertion of 得 (DE_{pot}) leads to that conclusion. He also points out that such constructions may be frozen at different levels, ranging from free combinations such as 打碎 (dǎ-suì) 'hit-break into pieces' to collocations such as 愛上 (ài-shàng) 'to fall in love' (*lit. love-entering.*

However he does not use the insertion test as defined in previous works but a more detailed version specifically formulated for dependency grammars, which he calls the *separability of a linguistic sign* (see below). He rejects the insertion test but unlike Duanmu, he argues that it can only be used to characterise phrases. If the insertion is impossible, it cannot be concluded that the expression is necessarily a word.

[4]Considering these expression as compound is questionable. However they may be treated as "words with spaces" as defined by Sag et al. (2002), see chapter 4

3.2.10. Intuition

Duanmu mentions the use of native speaker intuition, but only to reject it:

> "On the other hand, the fact that there is still no consensus on where to draw the line between word and phrase in Chinese, even though the discussions started since at least the 1950's, indicates that there are areas where people's intuitions either are not clear or do not agree."

In our work, we avoid relying on intuition for both theoretical and practical reasons.

3.2.11. Applying the Criteria

We will now discuss how these tests are used. To decide between words and phrases, N'Guyen relies on Mel'čuk's definition of *autonomy* which can be either *strong* or *weak* Mel'čuk (1994). If the autonomy of each constituent of an expression is sufficient, then the expression is a phrase. If the constituents are non-autonomous, then the expression is a word. The difficulty arises when it is necessary to define what constitutes a "sufficient" autonomy.

Strong autonomy is defined as follow:

> "a segmental sign X has a strong autonomy in a language $\mathscr{L}$ if and only if there exists a complete utterance in $\mathscr{L}$ which contains X and no other segmental sign"

In addition to the classic tests discussed above, N'Guyen proposes three criteria, adapted from Mel'čuk (1994). These three *first level critera* allow him to discriminate between three statuses:

- a strong autonomy, in which case the expression is clearly a phrase
- no autonomy, in which case the expression is clearly a wordform
- a weak autonomy, in which case *second level criteria* must be applied.

The three *first level criteria* are as follows:

1 - Separability of a linguistic sign

> " A segmental sign X is considered separable in a context XΨ or ΨX if and only if it is possible to insert an expression made of autonomous signs between X and Ψ without changing either their relative position, their semantic relation or their respective semantic content".
> (Mel'čuk,1994) translation is our own

Here XΨ or ΨX has a strong autonomy and Ψ is autonomous (weakly or strongly).

If an expression is proven separable, it follows that its parts are sufficiently autonomous. Thus the expression is treated as a phrase. If an expression is not separable, no decision can be taken and we must apply other tests.

2 - Distributional variability

> "A segmental sign X is considered distributionally variable if and only if Ψ belongs to more than one syntactic distributional class and neither the semantic relation between X and Ψ nor their respective semantic content depends on the class of Ψ"
> (Mel'čuk,1994) translation is our own

3 - Transmutability A segmental sign X is defined as transmutable in a context XΨ or ΨX if X and Ψ can be permuted or if X can be transferred to an other host. This property is typical of phrasal affixes and clitics.

Duanmu provides guidelines that are less formal but mostly consistent with N'Guyen's proposal.

> "The assumption here is that phrases should have a regular syntactic and semantic behavior; they should allow conjunction reduction, be made of free parts, be semantically compositional, and be structurally endocentric. If an expression fails any of these tests, it is not a phrase. This assumption is held by all analysts and will not be disputed here."(*sic*)

We do not subscribe to this assumption. It supposes that syntax is strictly regular and that its units should have both a clear semantic contribution and to clearly belong to a distributional class. It may be the case for quantity of phrases, but counter examples are pervasive in any language if one considers the Multi-Word

Expressions (see chapter 4). Under this assumption, wordhood criteria can be seen as heuristics to deal with the cases that are unclear or contradict the assumption. We think that these various criteria have more potential and that it is more interesting not to consider the difficult cases as exceptions.

We also see no reason why syntactic units and semantic units should be expected to match.

Huang (1984) argues that most of the relevant tests can be derived from the *"Lexical Integrity Hypothesis"* (LIH) (Chomsky, 1970; Bresnan, 1982) which states that "*No phrase-level rule may affect a proper subpart of a word.*" This hypothesis highlights the need for a comprehensive description of the grammar and its phrase level rules to be able to assess for the wordhood of a form.

3.3. Minimal Units in Syntax and Semantics: The Case for MSC

In one of our own previous works (Magistry, 2008), we started from the formalisation of the minimal units proposed by Kahane (2008) for French and adapted his definitions to MSC. The strength of Kahane's proposal is to distinguish clearly between minimal units of form, of meaning and of syntax and to acknowledge from the start that there is no perfect overlap of these three levels. Although we went through this adaptation without any major issues, this formalisation suffers from a drawback common to methodology presented previously: The need for certain grammatical knowledge prior to segmentation.

Kahane's "words" correspond to a specific subset of *syntaxemes* with the properties of *indissociability* and *weak autonomy*. We will now provide a definition of these terms with examples in MSC.

Morphemes (or *monemes* to follow Martinet's (1960) terminology) are the minimal units of form. They are defined as "a maximal collection of minimal linguistic signs of related meaning and form." They may not be easy to delimit in French but in written MSC they correspond well to the Chinese character.[5] A *mone* is then

[5]This is an approximation: counter examples are i) completely bound characters (under DeFrancis's 1984 terms, 蝴蝶 húdié should be considered as a single morpheme rather than a

defined as a sub-collection of signs from a *moneme* that have compatible distributions.

Let us consider the Chinese character 家 jiā as an example. All its uses share the denotation of *house, family* or *group of people*. They can thus be grouped together as a common morpheme. However we can distinguish between four different usages (or related distributions) for this morpheme:

- the noun *family, house, household*;
- a nominal classifier for companies;
- a nominal suffix to build the name of an expert of a field from the name of the field (often translated into *-ist*);
- the second (bound) character in the noun 專家 zhuanjiā *expert*.

These four distributions define four distinct *mones* of the same morpheme.

Syntaxemes can be made up of a single morpheme or a sequence of several morphemes (called a *polymone*). To be a syntaxeme, a mone must have a defined syntactic distribution and thus belong to a distributional class. This is the case for 家 jiā 'family' but not for 家 jiā the second character from the word 'expert'.

Kahane defines the *indissociability* of distributional classes as follows:

> Two distributional classes are **indissociable** if a member of one can never appear without a member of the other. Two *syntaxemes* are **indissociable** if they belong to distributional classes that are indissociable from each other. A sign is **indissociable** if all its possible decompositions lead to indissociable parts.

This property is important to account for inflection and thus less relevant for MSC, but we may argue that it applies for nominal classifiers that are indissociable from numerals or demonstratives.

With these definitions, we can distinguish

- **lexemes**, which are syntaxemes belonging to an open distributional class

polymone) and ii) there are a few examples of characters with multiple apparently unrelated meanings (they often have different readings), which should not be grouped together in the same *morpheme*.

- **inflectional morphemes**, which are syntaxemes belonging to a closed distributional class and are indissociable. This includes for example the aspectual suffixes in MSC, which are a closed class of three items indissociable from the class of verbs
- **grammatical lexemes**, which are syntaxemes that belong to a closed class and are not indissociable from an open class. This includes for example the prepositions.

Concerning the detection of minimally meaningful units, Kahane follows Martinet (1960) who suggests that a contrastive choice underlies every unit. Semantic units can be expressed either by lexemes or by phrases in the case of a locution.

These definitions do not distinguish between clitics and phrasal affixes. However Kahane presents two borderline cases of indissociability in French: verbal clitics and nominal determiners. Similar cases can be found in MSC. The aforementioned potential clitics 得 (de) and 不 (bu) are indissociable from the class of verbs, but verbs do not require such clitics. Another borderline case can be found in 以前 (yǐqián) and 以後 (yǐhòu). Liu and Oakden (2013) analyse these two forms as phrasal affixes.[6] As phrasal affixes, they require a phrase to attach to, but no distributional class strictly requires such an affix.

It seems to us that a non-symmetric definition of indissociability may be needed to describe phrasal affixes and clitics properly. But this is not a major issue for our purpose and we have to leave this question unanswered.

What matters most for the present work is that the definitions provided by Kahane (2008) rely on pre-supposed *distributional classes* for syntaxemes and on the ability to judge the "distributional compatibility" of morphemes. This drawback is shared by all the works presented with the exception of Kratochvíl who merely sketches out a methodology to acquire such distributional classes. This is a real obstacle issue as we reach a circular definition where segmentation is required to describe distributional classes which are in turn required in order to detect segmentation units.

In our previous work (Magistry, 2008) on quantitative morphology, we relied on a manual segmentation and part-of-speech tagging carried out by linguistic experts

[6] She uses the term "clitics" in her paper but her definition of this kind of clitic corresponds to what we call phrasal affixes

(we worked on the Sinica Corpus, see chapter 6). This was a major limitation of our work as it required a large quantity of language data and manual annotation can be intractable or costly. We could not scale up or try our system on new data.

On top of that, the reliability of automatic annotation for linguistic studies is questionable (this will be discussed in chapter 6).

At this point, we have found no methodology in the literature that would enable the computability of a linguistically sound definition of wordhood.

3.4. Computability of Wordhood

We will not go further into the details of the application of the previous tests by humans. Our concern is that they are simply not applicable as they stand for an automatic processing of a large body of texts. The use of these methodologies requires some prior knowledge about *word classes* and syntactic rules. This can in fact be considered as an inductive process rather than a deductive procedure where tests are used to falsify or validate hypothesis about wordhood within a specific description of a grammar of MSC. The main underlying idea is that the *word* in Chinese linguistic literature is generally considered as a minimal syntactic unit. Depending on the linguistic framework at hand, it is described as an element that can occupy an X^0 position (X-Bar theory), a vertex of a surface dependency tree graph (MTT) or an atomic element for a c-structure in LFG (Kaplan and Bresnan (1982)). In any case, a great deal of prior knowledge is required to initiate the induction.

Differences in the chosen linguistic framework can result in different conclusions about wordhood. For example, as we mentioned in example (12), N'Guyen regards Resultative-Verb constructions such as 看 見 (kàn jiàn) *see (look perceive)* as phrases because it is possible to insert a clitic to have the potential form. Duanmu on the other hand follows Chao and decides to treat 看見 (kàn-jiàn) and 看不見 (kàn-bu-jiàn) *unable to see (look-not-perceive)* as compound verb forms. We believe that this distinction arises from the fact that the Meaning Text Theory provides a specific way to deal with *collocations* and pays close attention to *"multi word expressions"*. We will come back to the relationship between Chinese Word Segmentation and Multi Word Expressions in Chapter 4. Here we just stress the influence of the linguistic framework on word segmentation. Syntax manipulates words but words seem to emerge from the syntactic description.

Many tests rely on word classes. For example, the productivity criterion requires parts of speech. If we want to say that unlike in English, $NP \rightarrow A\ \ N$ is not productive

in MSC, we have to define what we mean by *A* or *N*, and if the rule is not productive, it may be that there is no distributional category that can be called *A*. Such information is usually missing in the word segmentation literature or presented as something intuitive.

Kratochvíl points out the danger of relying too much on intuitions that are likely to be driven more by academic knowledge of the etymology of MSC and Classical Chinese than actual synchronic language data.

> "It is very current among sinologues to speak of 'nouns', 'verbs', etc., when they refer to forms which in fact occur only as root morphemes in MSC, and it is not uncommon to find even MSC affixes labelled that way. Comparaison betwen old Chinese and MSC in this respect is of unquestionable importance for historical linguistics, but mixing units of various historical levels of the language often makes the discussion on MSC word classes even more confused."

He also provides some direction towards a systematic methodology of the definition of *word classes* in a purely synchronic way, but this has never been systematically developed into a large scale description of the MSC lexicon, which is not so surprising given how such a task would be time consuming if addressed manually :

> "Another reason why the form class membership hypothesis has never been put through a sufficient test is the very practical point that such testing would be extremely time consuming and complicated."

He was also concerned by the circularity of the analyses which rely on deeper linguistic structures such as syntax and semantic representations and requires the selection of a linguistic theory, including the specific description of MSC under that theory, before looking at the actual data.

> "This circularity of defining syntactic functions by word classes and vice versa is perhaps an unfortunate result brought about by the obvious need to find a foothold in dealing with the structure of a language which lacks the familiar formal apparatus and in which everything consequently seems vague and evasive.
>
> What MSC linguistics needs are not exercises in logic or brief sample displays of techniques favoured by this or that general linguistic school, but a presumptionless large-scale analysis of MSC syntax which would,

> beside others, establish word classes in terms of form class membership. It is the author's belief that such an analysis is feasible."

In our work, we do nothing more than pursuing the same goal of an agnostic exploration of the language data. The only difference is the computational power now available. We have a new tool to conduct the exploration and analysis of the data which was not available at the time Kratochvíl was writing. It enables us to process the data at an unprecedented speed, but it comes with its own specific challenges and limitations.

A large scale analysis under a specific theory would necessarily include every levels of linguistic analysis. Doing so is not totally unthinkable but to be fair, we would have to allow the iterative modification of our grammatical description as well as the theoretical framework itself as we cover more and more phenomena. This probably has to be done manually and is practically intractable. Thus we choose to proceed with an agnostic yet mechanised exploration of the surface data.

On the other hand, we have to lower our goals to be able to reach a high degree of mechanisation. We will also face specific challenges that arise when language intuition is unavailable. These are the limitations that we will discuss in chapter 5. We will see in chapter 6 that we are still aiming for a linguistically more relevant analysis than what is typically done in NLP. Before this, we will see in the next chapter that the difficulties to define the minimal processing units goes beyond the word-level and extends with striking similarities to the so-called "multi word expressions."

Chapter 4

Beyond Chinese: Word Segmentation and Multi-Word Expressions

When one tries to connect the problem of Chinese Word Segmentation (CWS) to general linguistic and NLP of other languages, a striking fact is that it is formally and computationally closely related to what is called "Multi Word Expressions" (MWE) for languages using other writing systems.

The community of researchers involved in CWS is trying to separate *words* from one another, whereas works about MWEs try to relate *orthographic words* that should not have been tokenized separately but should be grouped as one word. Both are doing so to find more relevant units for further processing or to build lexicons. Surprisingly enough, they share many linguistic criteria as well as algorithms but are rather disconnected in terms of academic fields.

The following example will underline the similarities.

(1) *une peur bleue*
a fear blue
'a huge fear'

Although both *peur* and *bleue* are free in other contexts,

- the meaning of *peur bleue* is not compositional, *peur* does have its usual semantic contribution but *bleue* does not.

- the structure of this NP is endocentric (*peur* can be considered as the head)
- *bleue* lost its ability to accept modifiers such as adverbs: * *une peur très bleue* (*"a very blue fear"*)
- the rule NP → Det N Adj is productive but *peur* cannot be modified by an other color than *bleue*.

On top of that, proposed analysis from distinct theories may not agree. In MTT, *peur bleue* will be treated as a phrase were *bleue* is a collocative for *peur*, modelled using the *Lexical Function* ***Magn***(), as an usual way to intensify the meaning of *peur*. On the other hand, the *Tables du Lexique-Grammaire* (Gross, 1975) regards *peur bleue* as a compound noun. This situation resembles in many ways to the example of 看見 (kànjiàn) *look-perceive* seen in the previous chapter.

It seems to us that works towards CWS and MWE have a lot in common, especially the fact that they are both questioning the relation of lexicon to morphology and syntax.

4.1. Lexicalism in Linguistic Description

Huang (1984) claims that the *Lexical Integrity Hypothesis* (LIH) underlies most of the tests we presented in Chapter 3.

> "No phrase-level rule may affect a proper subpart of a word."

As a matter of fact, all the analyses we are aware of are done under this hypothesis for Chinese Word Segmentation.

The origin of the LIH is traced back to (Chomsky, 1970) and is extended in (Bresnan, 1982). Chomsky posits that the syntax must not account for the irregularities observed in constructional morphology (taking example from nominalisations in English). Bresnan extends this observations to flexional morphology. It follows the independence of the Syntax and the Morphology, the latter being responsible for the creation of the lexicon where lie all the irregular forms and the Syntax is implicitly supposed strictly regular.

Duanmu (1998) adopts this stance when he states that "*The assumption here is that phrases should have regular syntactic and semantic behavior; they should allow conjunction reduction, be made of free parts, be semantically compositional, and be structurally endocentric. (…)This assumption is held by all analysts.* (sic)"

The point made by Chomsky and Bresnan is that there must exist a lexicon independent from the syntax. The forms contained in the lexicon were generated by morphological rules (derivational, flexional or constructional) whose both input and output may be element of the lexicon. The syntax then uses the forms readily available in the lexicon. The syntax does not have the capacity to modify the lexical items and can only combine them regularly.

Note that Although N'Guyen claims (Nguyen, 2006, p.77) that he is consistent with the LIH when he states that no part of a wordform can receive a modifier, we may argue that as a lexical description in MTT includes information about the collocations and their "*Lexical Functions*", in MTT the lexicon and the syntax are not strictly autonomous.

However, it seems to us that syntax is also responsible for the creation of frozen idiomatic expressions with regular internal structure that can in turn be *borrowed* into the lexicon and used in morphology to create new lexical items. We use the term *borrowed* because this process resembles to the borrowing from another language's lexicon (Walther and Sagot, 2011). Derivation rules can then create exocentric items like *un je-ne-sais-quoi* in French. What was once a regular syntactic construction thus becomes irregular or opaque through linguistic change. This fact has been neglected by early works in formal linguistics even though it was described as early as in (Jespersen, 1924). We think that this is particularly important for Chinese as the lack of inflectional morphology may make compounds difficult to distinguish from frozen idioms which became lexemes.

More recently, the strict regularity of syntax came under question in at least two ways: the gradual aspects in possible alternation between competing rules (Bresnan, 2007; Bresnan et al., 2007; Thuilier, 2012) and the integration of Multi-Word Expressions to the lexicon. Psycho-linguistic experiments show that the mental lexicon include both opaque and transparent MWEs for which a certain frequency threshold has been reached (Sosa and MacFarlane, 2002). Lower estimates for the

part of MWE in the mental lexicon are of the same magnitude as simple word expressions (Church, 2013; Jackendoff, 1997).

We consider that syntactic freezing and borrowing is an important operation for constructional morphology. It must be accounted for, along with derivational and inflexional morphology when we describe the processes of lexeme creation.

Since MSC has very poor inflexional morphology, we cannot include inflexional paradigms in the description of its lexicon. Constructional morphology is thus responsible for a large majority of lexemes in the MSC lexicon and the distinction between compounding, derivation and *borrowing from syntax* becomes more fuzzy. As a result, the need to account for syntactic freezes is more obvious in MSC, Hsu (2012) even claims that the decision to treat some expressions as dynamically created phrases or frozen lexicon items can evolve on a very short laps of time. We regard this phenomenon as a graded one.

4.2. Jespersen's opposition

Jespersen (1924) distinguishes *formulas* from *free expressions*. As an example of the latter, he argues that *"John gave Mary the apple"* and *"My uncle lent the joiner five shillings"* share a same abstract and productive structure. This contrasts with *"Long live the King !"* whose structure cannot be used to form new utterances like *"*Soon come the train !"*, or *"*Late die the King"*.

Jesperson notes that:

> "The distribution between formulas and free expressions pervades all part of grammar"

He also stressed the interaction between formulas, language evolution and written norm. He compares formulas at the utterance level to irregular morphology. In English, the once productive suffix *-en* for the plural has only one resulting form left (*ox/oxen*) in the lexicon. Forms like *shoen, eyen* have been reshaped into *shoes and eyes* and *-en* has been supplanted by *-s* as the only productive mark for the plural of newly coined words.

More importantly for our discussion, he argues that what can be said about the productivity of word-formative affixes such as *-th* and *-ness* and the subsistence

of older forms as frozen formulas holds at a syntactic level, for example within compounds. He gives example based on *hūs (house)* such as *husband, hustings, hussy* which all refer to *house* and were initially formed as regular syntactic phrases. As those compounds underwent regular sound changes and as the orthography evolved the connection between these three words became imperceptible without the knowledge of their etymology.

4.3. MWE in Lexicography and NLP

The question of MWEs have been addressed mostly by lexicographers, for example in COBUILD and DELAC (Courtois et al., 1997; Gross, 1996, 1986; Sinclair, 1987). More recently, they received more attention in NLP as a proper treatment of MWEs has been proven useful to improve parsing (Constant et al., 2011; Sag et al., 2002; Green et al., 2011).

From a lexicographic point of view, MWEs present specific difficulties for non-native learners. In French we can also observe discrepancies in orthography among native speakers and even among dictionaries (Mathieu-Colas, 1988) for both punctuation and inflection of MWEs.

What distinguishes compounds and simple words in the DELAC is the punctuation. This may result in a different treatment of orthographic variants for a given lexeme. For example, *delta-plane* is considered as a compound word but *deltaplane* is a simple word.

In French, the inflection of compound words can also be particularly tricky and results in inconsistencies or disagreement among native speakers. These disagreement can even be the object of spelling reforms. Consider for example the words *betterave*'beetroot' and *chou-rave*'kohlrabi'. Both words in French are built on the same pattern *X+rave*'X+turnip' with a compositional meaning. The higher frequencies (or degree of familiarity) of *betterave* over *chou-rave* and of *chou* over *bette* may explain a higher degree of opacity for *betterave* among French speakers and the differences in orthography. Interestingly this result in two different ways to mark the plural forms:

(2) a. des betterave**s**

b. * des bette**s**rave**s**
c. des chou**x**-rave**s**
d. ??? des chou-rave**s**

This kind of orthographic peculiarity can even become the subject of political reforms.

As MSC does not mark the plurals. Different insights are required to treat 甜菜 ∘ 根 ‘beetroot’.
tiáncài gēn

In NLP a widely accepted classification of MWE was proposed in (Sag et al., 2002). Where the authors first distinguished *Lexicalized phrases* from *Institutionalized phrases*. They define the latter as formed by regular syntactic rules but being preferred over any possible paraphrase. The former being irregular in some way that make them more or less frozen. *Lexicalized phrases* are further divided into *fixed expressions*, *semi-fixed expressions* and *syntactically flexible expressions*, in decreasing order of lexical rigidity.

Fixed Expressions are immutable expressions that include spaces but do not allow for any kind of modification or insertion. They include some proper names such as *Palo Alto*, Latin idioms like *ad nauseam* but also common English expressions “that defy conventions of grammar and compositional interpretation.” (e.g. *by and large, in short, kingdom come, and every which way*). Sag et al suggest that such expressions are fully lexicalized and are thus to be represented as “word-with-spaces”

Semi-fixed Expressions include nominal compounds and non-decomposable idioms. They share the property of being syntactically inalterable (their parts cannot receive modifiers and they cannot undergo syntactic variation like passivization) but the inner parts may be inflected. Examples are *attorney general* which inflect into *attorney**s** general* and not **[attorney general]**s***, and *shoot the breeze* where the verb can be inflected but the object cannot be modified. The authors argue that **semantic non-decomposability** is responsible for the impossibility to alter the syntactic structure of such idioms. However the decision about syntactic immutability can be tricky to make. The example chosen by the authors is presented as impossible to use in passive form: Sag et al. judge *“*the breeze was shot”* unacceptable. But not every speakers would agree as it can be found in some context: *“So more alcohol was*

consumed and the breeze was shot."[1]

Syntactically Flexible Expressions include idioms that are to some extent **semantically decomposable**, for example the idiom *let the cat out of the bag* has parts that can be modified as in *"The cat is **now** out of **NSA's** bag"*.
In this category, Sag et al. also include the verb-particule constructions that are very common in English an may present idiosyncratic meaning, and impose different constraints regarding insertion of adverbs or nominal object between the verb and the particle. Light verb constructions such as *make a mistake* or *give a demo* are also included in this category.

Institutionalized Phrases are regularly formed phrases. Their parts can be inflected and modified following the usual and productive rules of the grammar. They are also semantically compositional. However, they appear at an unexpected high frequency and are largely favoured over possible paraphrases. Paraphrases may even seems unnatural to native speaker or the choice of one paraphrase over another may be indicative of a local dialect or sociolect. Examples include *telephone booth* which is specific to American English and can be opposed to *telephone box* in other variants of English or *???telephone cabinet* which could be understood and is regularly formed but appears unnatural.

This categorization gives a good overview of the variety of phenomena and relates lexical rigidity to the relevant morphological and syntactical aspects. We may however want to complete the picture and include Jespersen's discussion about compound words without spaces whose inner parts may be more or less opaque. We would end up with a scale schematized on Figure 4.1

4.4. MWEs and the Dissociation of Syntactic and Semantic Units

In section 3.3, we suggested that a source of confusion to define the word in Chinese is the assumption that syntactic units and semantic units should match.

MWEs constitute a clear case of mismatch between these two levels of analysis. If we follow Martinet's 1960 suggestion to rely on the successive choices a speaker has

[1] `http://manc_ill_kid.blogspot.fr/2006/10/sunday-i-was-supposed-to-be-recording.html`, (blog's author declares to be living in London)

	multiple morphemes	includes space(s)	internal flexion	allows for modifiers	compositional meaning
Simple Word	No	No	No	No	No
Compound Word	Yes	No	No	No	No
Fixed Expr.	Yes	Yes	No	No	No
Semi-fixed Expr.	Yes	Yes	Yes	No	No
Syntactically flexible	Yes	Yes	Yes	Yes	No
Phrase	Yes	Yes	Yes	Yes	Yes

Table 4.1.: **types d'expressions multi-mots et critères utilisés dans cite sag et al 2002**

to make to build an utterance, the use of an expression like *kick the bucket* is more likely to result from a single choice rather than a choice to talk about a *bucket* of which one chooses to say that it is *kicked* by someone. In this case we would conclude to a single semantic unit conveyed by a phrase, that is multiple syntactic units. This situation is problematic only if we want to keep the assumption that the two levels should match. It seems to us that MWEs constitute a strong argument to reject it.

Examples of *fixed expressions* given in (Sag et al., 2002) can all be described as a single syntactic unit (in terms of distributional classes). From our point of view, they are *polymones*. We do not consider the fact that the orthography of a *polymone* includes a space as relevant. When an inner part of an expression can be inflected, it follows that we can attribute a word classe to the subparts. Therefore Sag et al.'s *semi-fixed expressions* are made of multiple syntactic units but they results from a single choice of the speaker. The same goes for *syntactically flexible expressions*. The latter and *institutionalized phrases* are semantically decomposable. Nevertheless, both result from a single choice (we do not talk about a *booth* that happens to include a *telephone*, we choose to talk about a *telephone booth* "directly").[2]

[2]Although it may be hard to prove without psycho-linguistic data, for example about lexical access time.

4.5. MWE with Regard to Chinese Word Segmentation

The Chinese script is characterized by a surprising graphical stability. Throughout history, Chinese scholars and rulers have been more preoccupied by the "correct" readings of characters than by their actual adequacy with spoken languages (one reason for that being the fact that Chinese script was primarily used to write classical Chinese and "correct" is to be understood as "like in ancient times"). The result is a greater transparency of old compounds and etymology. This contrasts with languages such as English or French for which writing evolution tends to obfuscate the origin of words in order to follow their actual pronunciation. It is very likely that had he worked on Mandarin, the aforementioned compounds discussed by Jespersen (*hustings*, *hussy*, *husband*) would have been more transparent, not that Mandarin Chinese is not subject to phonetic evolution, but the writing form did not follow the spoken form at the same pace as it did in English. It is worth mentioning that English and French orthographies are not following closely the evolution of the spoken form either (compared for example to Italian or German) but Chinese is probably the more conservative writing on this scale.[3]

Now if we go back to Figure 4.1 with MSC in mind, we note that the first distinction between simple words and polymones is more transparent in MSC and fuzzier in English. There are however examples of multi-character words in Chinese that are in fact monomorphemic and should be regarded as simple words.

The second step (allowing spaces between roots) between *complex words* and *fixed expressions* is irrelevant for Chinese script and may be subject to discrepancies for languages written with in Latin script. Both types can be considered as polymones, the distinction between the two may thus be irrelevant.

The third step is also not very relevant in the case of MSC because of its lack of inflectional classes, *semi-fixed expressions* may thus be delicate to distinguish from *fixed expressions* in MSC. It may only concern verbs if we treat aspectual and potential markers as suffixes. If we don't, insertion of an aspectual marker will be considered as a modifier of the verb and the expression will thus be *syntactically*

[3]Chinese is far from being an ideographic script, a large majority of the Chinese characters were created on a phonological basis (over 90%). But most of them were accounting for aspects of Archaic Chinese phonology (Sagart, 2006). The etymology of characters is thus easier to trace back, and with it some inherited semantic relatedness at a morphological level.

flexible.

Allowing the insertion of modifiers in some parts of an expression has been described as a clear indication of a multi words phrase (it is a case of *XP substitution*). The same goes for syntactic alterations. It can be argued that there is no passivization in MSC that would be strictly equivalent to passivization in English so we don't consider this test, but we can mention other structures such as Yes/No questions and 連 (liàn) X 也/都 (yě/dōu) Y constructions that yield similar results in terms of wordhood and segmentation tests, as in example (3-b) quoted from (Chen, 1957) by (Paris, 1979).

(3) a. 人們 似乎 忘記 了 肚子 餓
rénmen sìhū wàngjì -le dùzi è
people as if forget ASP stomach hungry
'it was as if the people forgot that they were hungry.'

b. 人們 似乎 連 肚子 餓 都 忘記 了
rénmen sìhū lián dùzi è dōu wàngjì -le
people as if even stomach hungry all forget ASP
'It was as if the people had even forgotten that they were hungry.'

As far as segmentation is concerned, *Institutionalized Phrases* should be considered as phrases.

The different type of units and their properties in term of autonomy and number of orthographic words are summarized on Table 4.2. It seems to us that *wordhood* is fuzzy all the way from totally bound forms to regularly formed phrases. However, to address the the question of Word Segmentation, there seems to be a more salient cut between *semi-fixed expressions* and *syntactically flexible expressions.* as that is where syntactic rules come into play and multiple strongly autonomous units are combined. Even though for some expressions, the non-flexibility may be hard to prove.

This defines our goal in terms of word segmentation and may have consequences if we want to adapt the present work to other languages.

It is worth mentioning that, although we consider that such a fine-grain definition

type of unit	autonomy	# orthographic word(s)		# Syntactic unit(s)
		Latin[4]	Chinese	
Bound Root	No	<1	≥1	≤ 1
Affix	No	<1	1	< 1
Clitic	weak	1	1	1
Phrasal Affix	weak	1	≥1	?1
Grammatical Word	weak	1	≥1	1
word	≥weak	1	≥1	1
Fixed Expression	strong (one)	≥2	≥2	1
Semi-Fixed Expr.	strong (one[5])	≥2	≥2	> 1
Syntactically Flexible Expression	strong (many)	≥2	≥2	> 1
Institutionalized Phrase	strong (many)	≥2	≥2	> 1
Phrase	strong (many)	≥2	≥2	≥ 1

Table 4.2.: **Summary of all mentioned units**

of the different morpho-syntactic units is needed for linguistic description, it is not always a requirement for more practical NLP application. For example in the case of lexicon extension using specialized domain terminology extraction, Patin (2013) prefers to skip the tokenisation step to aim directly at meaningful units and avoid the propagation of segmentation errors to further processing steps.

[4]With the exception of romanised Vietnamese where the values of the Chinese column apply.

[5]Strictly speaking, it would be one strongly autonomous unit in MSC but non-autonomous inflection may be included in morphologically richer languages. This difference may affect segmentation objectives but in this dissertation, we focus on MSC.

Chapter 5

Towards a Corpus-based Definition of Wordhood

The previous chapters underlined multiple issues to define wordhood. The need for a detailed description of the grammar under a given linguistic framework makes it a chicken-and-egg problem. The minimal units of various levels of analysis can mismatch, for example a frozen expression can have a single non-decomposable meaning denoted by multiple syntactic units in an apparently regular construction. We think that it shall not be assumed that a canonical word is a syntactic unit with a specific meaning and that divergent cases are to be treated as exceptions. Quite the opposite, we do not expect the two levels to match. Nevertheless, we think that syntactic and semantic structures influence the distributions of forms in a corpus in a noticeable way that should enable us to break the chicken-and-egg problem.

In our work, we consider the unannotated (*raw*) corpus as the object of computational linguistics analysis. It contains all of our *observations*. For practical reasons and for the sake of a quantitative evaluation that would be comparable with previous works, we will re-use a variety of corpora that have been used for Chinese Word Segmentation tasks in NLP. We will assume that they represent some consistent state or sociolect of MSC.

We also claim to be agnostic in term of linguistic theory. By relying solely on raw corpus data, we indeed prevent ourselves from using the variety of information sources that are required for traditional theories to be applied. This includes speaker intuition but also syntactic analysis, semantic, phonology as well as experimental elicitation. We are left with a long sequence of characters and punctuation marks.

Our guideline is simple: any knowledge about an open set of items shall be induced from the raw data by computation.

The main hypothesis underlying our whole work is that language is structured. Thus language data must be distinct from a random string of characters in a way or another. The structure we are looking for shall emerge from this non-randomness of the data. The underlying structure is what allows humans to acquire a language and use it to communicate. The very fact that all the aforementioned criteria apply seems enough to believe that the same underlying phenomena that make the criteria useful will create biases in the data we observe in a corpus. These biases will serve as a foothold to start the induction of the segmentation and the building of the lexicon.
This hypothesis is essentially in line with the starting point of structural linguistics, all we do is to look for a way to mechanize the analysis.

The motivation for such a radical stance is two-fold. First of all we aim for a high degree of objectivity as our object and observations are clearly defined and delimited. The second one is more practical, we aim to lower the burden of manual labour required to segment text and classify wordforms.
The formal linguistics methodology presented in this chapter turns out to be impossible to fully automatize with the current level of formalization proposed by various linguistic frameworks. Therefore following such procedures can be impracticable if we aim for a large coverage of the lexicon. And we do.

However, we still build on previous linguistics works as we expect to face a variety of known phenomena. Especially the fact that wordhood is not a binary notion, we are expecting word formative and grammatical affixes as well as phrasal clitics. These have been well described for many lexical items so we can check whether the system we propose manages to capture them. We will also show that the classical methodology for defining word classes can be adapted to corpus linguistics.

5.1. Refining our Goals

We have seen in the previous chapter that the first criterion needed to define wordforms and other segmentation unit is their *autonomy*, also described as a *syntactic freedom*. The second kind of information is *word classes*. Those shall be derived from *substitution frames* but as Kratochvíl remarked:

> "The practical difficulty, of devising techniques for setting up word classes according to form class membership is caused by several factors mentioned above, primarily the lack of a sufficiently consistent criterion for identifying the borderlines of words."

Thus we shall first look for boundaries that we can rely on to detect *word classes*. Those boundaries are likely to be closely related to the *autonomy*, which makes defining an empirical measure for autonomy of an expression our main objective.

In fact, segmental autonomy and word class membership matches the two traditional axes of linguistic analysis: the syntagmatic axis and the paradigmatic axis, respectively.

The decision concerning the status of a given expression should be a combination of those two aspects : *autonomy* and *wordclass membership*. An expression should be considered a wordform if it has a high degree of *autonomy* and if it can be associated with a distributional *wordclass*. We will address the question of *autonomy* in the second part of this thesis. A methodology and preliminary experiments to define wordclasses will be sketched in the perspectives.

5.2. Corpus and Autonomy

In the previous chapters, we presented different definitions for the *autonomy* of an expression (which intuitively corresponds to its *syntactic freedom*). Unfortunately, nothing is strictly data-driven. They all rely on intuition, elicitation or deeper linguistic analysis.

The main issue we have to face is referred in corpus and computational linguistics as "*data sparsity*." As a matter of fact, the number of possible sentences in one language is infinite and our corpus is necessarily finite. The number of perfectly acceptable utterances absent from any corpus is therefore infinite. We cannot rely on elicitation and dynamically create (or ask a native speaker to do so) a new utterance that would fit our needs if possible. A corpus contains a limited number of sentences, and it may contains mistyping, genuine errors or data in a foreign language (from a quote or from code-switching) that would easily be discarded by a human analyst but not by a computer. As a result:

- not every possible production will be present in a corpus;

- not everything that is present in the corpus is to be considered as part of the language.

This means that we cannot rely on the fact that we cannot find a form in a corpus to decide that this form is invalid. On the other hand, corpora are very likely to contain noisy data to which we do not want to give too much credit. A form found in a corpus could be erroneous in many ways.

It follows that we will be unable to reach certainty. We will have to work with different levels of confidence. This is a typical use-case of a probabilistic modeling.

Note that we use only surfacic data without syntactic analysis. Therefore we do not strictly speaking measure the *syntactic autonomy*. These however are theoretical abstractions. What we aim to measure is the observable effect of the syntactic freedom on the distribution of surfacic data. In that respect it is not equivalent to the *autonomy* as defined in MTT but can be considered as an approximation for it. In the remaining of this dissertation, we use the word *autonomy* for both notions.

Not only the size of the corpus is necessarily limited, it is well known (Zipf, 1949; Baayen, 2001) that the distribution of the forms in a corpus is typically from the family of *LNRE* distributions (where *LNRE* stands for "Large Number of Rare Events"). This family includes the Zipf's law and the power law. What the distributions of this family have in common is that few items in the data are very frequent but most of the data consists of large quantity of unfrequent items. For language data, this means that some words are very frequent but most of the words present in a corpus will be rare and occur only a few times.

Let us re-use the Kratochvíl examples, and test the *reduction* procedure he proposes while replacing elicitation by lookup in a corpus. If we encounter the sequence 我們 wǒmen 也 yě 去 qù *we('ll) go too* in our corpus. It is possible, and even very likely that 去 qù *go* will not appear alone, as an isolated utterance in the same corpus. The risk here is to faultily conclude that as we cannot find it alone, we cannot operate a *reduction*. Therefore 去 qù would not be considered *autonomous*. This reasoning would be flawed in the sense that "absent from the corpus" is not equivalent to "agrammatical."

5.3. Corpus and Wordclasses

The wordclasses we need to capture are distributional classes that should group together forms that can appear in similar contexts. They correspond to those

described by Kratochvíl and are expected to be roughly equivalent to part-of-speech.

The main issue to define *wordclasses* is to select the relevant *substitution frames*. The three examples given by Kratochvíl are in fact very fortunate, and the way he filled the commutation paradigm is obviously guided by his intuition for the sake of illustration. If we ask a computer to do the same thing, we face a heavy combinatorial problem as the number of possible (yet useless) *substitution frames* will be extremely large. Considering all possible frames will be at best computationally very complex, and it will probably inject a lot of noise in our model. We illustrate the kind of noise that is to be expected in (1)

(1) 那 個 喜歡 她 的 人
nà *ge* *xǐhuan* *tā* *de* *rén*
This Cl. like her DE man
'the one who likes her'

Classifiers such as 個 (ge) are preceding the noun in a large proportion of noun phrases. We may thus want to consider the frame 個 *X* (*ge*) as a clue to classify X as a noun. However in our example, the presence of a relative clause places the verb 喜歡 (xǐhuan) *like* in the position X. We may want to leave out this kind of noisy data.

Intuitively, we can expect that verbs will be well caracterized by aspect markers, stative verbs will be very likely to occur with degree adverbials and nouns shall be found with nominal classifiers. This however require the knowledge of what is an aspect marker, an adverb of degree and a nominal classifier. In some cases like for aspects markers, the *substitution frame* can be characterized by an element belonging to a small closed set. It may be worth considering to manually include this information into our models. Another possibility is to follow Kratochvíl's insight : "*Grammatical affixes are thus direct word class indicators, since they are in a sense part of the frames on the basis of which form classes are established.*" This requires to induce what Kratochvíl calls "grammatical affixes," that is to say clitics, phrasal affixes and grammatical words, from the data.

5.4. Interaction Between two Graded Phenomena

We choose to rely on probabilistic modelling for two reasons. We already mentioned the first, for practical reasons we can only speak of degree certainty in an hypothesis given our corpus data. The second is that we have two objectives: *autonomy* and

wordclass membership are inherently two graded phenomena, and they interact with each other.

Let us illustrate our point with a small case study. It concerns a clear disagreement between different linguists on a set of examples :

(2) a. 人造 纖維
rénzào *xiānwéi*
man-made fiber
'man-made fiber'

b. 袖珍 詞典
xiùzhēn *cídiǎn*
pocket dictionary
'pocket dictionary'

c. 螺旋 推進器
luóxuán *tuījìnqì*
screw propeller
'screw propeller'

Those three examples are considered as two words each by Chao (1968) and also by Lü (1979). Duanmu cite those to reject the *syllable count criterion*, claiming that "the first immediate component is not a free form. (...) it conflicts with Freedom of Parts".

We now argue that this lack of freedom (or lack of *autonomy*) is relative but not decisive and we would favor the analysis in two words.

One thing is that we can perfectly insert a modifier in the middle of such expressions as in (3)

(3) 袖珍 法語 詞典
xiùzhēn *fǎyǔ* *cídiǎn*
pocket French dictionary
'pocket French dictionary'

which proves the two parts to be separable. This is clearly a blurry case in which many but not any items could be inserted. It may be argued that (3) is yet another wordform but that would create a family of related wordforms built on the same patterns (we can at least insert any name of language in this example). The productivity of this pattern may be subject to discussion.

Reduction cannot be applied to the three items and the shortest utterances we can form with the first parts include at least a clitic as in 人造 - 的 *(it is)man-made.*
rénzào de

Another argument is that the head noun in each of these three examples commute freely with large sets of strongly autonomous words. Under Kahane (2008) terms,

the firsts parts of a,b and c are thus defined as weakly autonomous.

We believe that what Duanmu considers a lack of freedom is also related to the difficulty to define the class membership of 袖珍 xiùzhēn *pocket*. He argues that it is neither a verb nor a noun. Adjectives in MSC are often claimed to be non-existent or to correspond to the class of stative verbs. But those three items are clearly not stative verbs either (they can't take any adverb of degree). Depending on whether we consider a small wordclass of adjective-like items which would include those three (and other words describing non-graded qualities), we will or not consider the pattern forming (2) as productive or not. If we do so, we will prefer a phrasal analysis. If we don't we will conclude they are compound words. Actually, they are very close to many cases of "*Multi Word Expressions*" in other languages such as French or English.

5.5. Towards Formalization

We can now make a step further to formalize what we are trying to do. To define the autonomy and wordclass membership of any expression present in our corpus, that is any sequence of contiguous characters, we will need to define :

- a measure of *autonomy* assessing the syntactic freedom of the sequence, it should have the form of a graded scale going from *bounded* to *strongly autonomous*.
- a procedure to induce *wordclasses* based on relevant *substitution frames*

Our proposition for the first point will be presented in details in Part 2 as our main contribution.

The second point implies the need for a segmentation procedure and a way to select relevant frames. We will show that our proposition for an *autonomy* measure can be used to perform segmentation. It should also help to define a procedure for *wordclass* induction as suggest our preliminary experiments in chapter 12.

We can now consider the aforementioned types of units under this newly defined objectives.

Words Words will be minimal autonomous sequences of characters to which we can attribute at least one *wordclass*.

Phrases Phrases are defined for a deeper level of analysis. Thus we are not looking for them explicitly. However they shall be segmented and we expect idiomatic expressions to be autonomous combinations of parts that may have a lower degree of autonomy. Ideally, we may want to consider 袖珍 xiùzhēn 詞典 cídiǎn *pocket dictionary* as a particularly autonomous phrase made out of two words, of which the first demonstrate a relatively low autonomy.

Affixes and clitics Given the unavailability of phonological and syntactical information, we may not be able to distinguish phrasal affixes, clitics and grammatical words. They are all forms bounded at a phrasal level. However if we define clitics in MSC on the neutralisation of the tone, this concerns only a very small and closed set of items for which specific treatment may be considered.

Affixes are expected to impose restrictions to their cooccurring sequences, this shall be reflected in their autonomy and this can be used to define *substitution frames*. Empirical observations about expressions such as 袖珍 xiùzhēn + 的 de *pocket+clitic* will be of special interest to judge the quality of our measure.

Chapter 6

Chinese Word Segmentation Bakeoffs and Resources for Automatic Processing

6.1. Introduction

As seen in chapter 3, the question of how words could (and whether they should) be segmented has been raised even before the beginning of the process of standardization of MSC. It is now clear that it is required for a concise and coherent description of the language. However, linguists are still arguing on how such a segmentation should be done. Work produced by linguists may focus on principles, such as those we presented in chapter 3, and other on specific lexicon items (Liu and Oakden, 2013). Nevertheless, NLP practitioners had to perform "Chinese Word Segmentation" since the very beginning of Chinese NLP and have produced a huge amount of work on that task in the last thirty years. In this chapter, we sketch the evolution of the field from the dictionary based approach to various machine-learning based systems. We will focus more on the work closers to ours: the unsupervised machine learning approaches. To make our discussion easier, the next session will present the evaluation methods and the relation between CWS in NLP and the underlying linguistics issues.

6.2. Evaluation Metrics

In quantitative evaluation of segmentation systems, wordhood and boundaries are binary notions. An expression is or is not a word and between two characters there is a boundary or there is not. The distinction between words, affixes and clitics is not accounted for at this level. There are thus two ways to score a segmentation: the first one is to evaluate the quality of the boundaries and the second one is to evaluate the quality of the segmented words. Both use the measure called *F-score*.

Here is how it is computed on words with an example, where we consider (1) as the reference and (2) as a candidate segmentation outputted by a system.

(1) 為何 會 有 一 群 人 自願 無償 寫 程式 服務
Why may have one group people voluntarily for-free write code serve
大衆
population ?
'how can there be a bunch of people willing to write code for free to help the population ?'

(2) 為何 會 有 一群人 自願 無償 寫 程式 服務
Why may have a-group-of-people voluntarily for-free write code serve
大 衆
big multitude

F-score is the harmonic mean of two other measures. The advantage of the harmonic mean over the geometric mean is that it penalises unbalanced results. Two medium values will yield a better f-score than a bad one and a good one.

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Word Precision is the ratio of words correctly segmented by the system to the total number of words in the output of the system. In this case, there are 12 words in the output of the system, 9 of which are correctly recognized.

$$P_w = \frac{\#\text{correct words}}{\#\text{words in the output}} = \frac{8}{11} \approx 0.73$$

Word Recall is the ratio of words correctly segmented by the system to the total number of words in the reference segmentation. In this case, there are 12 words in the reference 9 of which were found by the system.

$$R_w = \frac{\text{\#correct words}}{\text{\#words in the reference}} = \frac{8}{12} = 0.67$$

Word F-score is thus the harmonic mean of the two previously defined measures:

$$F_w = 2\frac{P_w R_w}{P_w + R_w}$$

For a boundary based score, we use similar measures but focus on the splitting decision, we illustrate these decisions on our example in (3) where we add dots (•) to figurate boundaries. The black dots are boundaries on which the reference corpus and the tested system agree. The green dots (•) are boundaries from the reference corpus that are missing in the system's output and the red dots (•) are boundaries not present in the reference corpus but added by the system.

(3) 為何 • 會 • 有 • 一•群•人 • 自願 • 無償 • 寫 •
Why may have a-group-of-people voluntarily for-free write
程式 • 服務 • 大 • 衆
code serve big multitude

In this case we can define:

Boundary Precision is the ratio of boundaries correctly set by the system (black dots) to the total number of boundaries in the output of the system (blacks and red dots). In this case, there are 10 boundaries of which 9 are corrects

$$P_b = \frac{\text{\#correct boundaries}}{\text{\#boundaries in the output}} = \frac{9}{10} = 0.9$$

Boundary Recall is the ratio of correct boundaries in the output of the system (black dots) to the total number of boundaries in the reference segmentation (black and green dots). In this case, there are 11 boundaries of which 9 were found.

$$R_b = \frac{\text{\#correct boundaries}}{\text{\#boundaries in the reference}} = \frac{9}{11} \approx 0.82$$

Boundary F-score is thus the harmonic mean of the two previously defined measures:

$$F_b = 2\frac{P_b R_b}{P_b + R_b}$$

Since an error on a boundary will affect the two adjacent words, boundary scores are higher than word scores for a given segmentation.

On top of these measure, an interesting value for supervised or lexicon-based system is the "OOV Word Recall". It gives the word Recall (R_{oov}) computed only for the words present in the corpus that were absent from the training data.

When training data is provided, a distinction is made between *closed* and *open tracks*. When the segmentation system make use of external data, such as in-house lexicon or other training corpora, it is considered competing on an *open track*. To compete on the *closed track*, one have to rely solely on the provided training data.

6.3. A Short History of Chinese Word Segmentation as an NLP Task

Trends in the development of Word Segmentation algorithm are closely related to the availability of language resources at a given time.

The turning point was the release of various manually segmented large corpora. The first one of this kind we are aware of was published in Taiwan by Academia Sinica in 1996 (Chen et al., 1996). Before that time, NLP practitioners could only rely on dictionaries, word lists or raw corpora to perform CWS. The main issue back then was two-folded: how to use a dictionary to segment a corpus and how to extend the dictionary to account for words that are present in the corpus but not in the dictionary. A noticeable exception is the work by Sproat and Shih (1990) which as far as we know is the first published work that relies only on statistics from raw corpus and is therefore the first work in unsupervised word segmentation.

The availability of the Sinica Corpus in 1996, followed by the Penn Chinese Treebank (Xia, 2000) and the Peking University corpus (Yu et al., 2002) was not only

the starting point of supervised machine learning, it made possible a comparable evaluation of the various proposed systems. Before that, different systems were evaluated against different home-made small corpora which made any comparison difficult.

After two segmentation contests held in China that were based on a single corpus (and thus a single segmentation guideline), a first international "Bakeoff Evaluation" was held at the second SIGHan Workshop at ACL in 2003 (Sproat and Emerson, 2003). During this bakeoff, all competing systems were evaluated against the same four reference corpora that are based on four distinct segmentation guidelines. More will be said about some of these corpora in the next section.

Twelve teams participated in the first bakeoff. Not every team provided results on every corpora. Systems word F-scores (F_w) ranged from 0.89 to 0.96 except for the corpus extracted from the Chinese Penn Treebank (CTB) which was the smallest and for which some inconsistencies in the annotations have been noted by participants (Results on the CTB fall between 0.73 and 0.91).

The more rigorous evaluation showed that previously published systems were probably overrated. It also pointed out the disastrous effect of data sparsity as OOVs seemed to be a major cause of errors (the best reported out of vocabulary word recall, R_{oov} is only 0.76 and most systems showed a much lower rate, especially on the closed tracks). Lexicon-based approaches are more likely to be affected by this issue so supervised machine learning took the lead and became the main research direction.

A second bakeoff was held in 2005 (Emerson, 2005). 34 systems were evaluated on four other corpora presented in the next section. Although some progress was made (overall, the best reported word F-score reached 0.97, and the median of the F_w of all competing systems was 0.94), OOV words were still a major issue with a best R_{oov} of 0.81 on the closed track. This indeed is worrisome as it reflects the inability of the algorithms to adapt to new data. It was to be expected that the performance of the competing systems would drop badly if used to segment texts from a different register or on too different topics.

Another major source of OOV are the Named Entities (NE), the third bakeoff held in 2006 (Levow, 2006) thus focused on both CWS and NE recognition. Although the results on the NE recognition task were fairly good, it didn't seem to improve much the treatment of OOV on the CWS task.

The fourth bakeoff (Jin and Chen, 2008) added a POS-tagging task but on gold-segmentation. For CWS task, the results were mostly consistent with the previous bakeoff.

The fifth segmentation bakeoff (Zhao and Liu, 2010) explored domain adaptation. The systems were tested on four different domains (literature, computer science, medicine and finance) but were provided training corpora from a fifth domain (journalistic). For litterature and computer domains, small unsegmented corpora were also provided and the participant knew they will be tested on these from the beginning. The two other domains (medicine and finance) were kept secret until the evaluation. Different corpora of comparable domains were made to test the systems on simplified and traditional Chinese.

The results show a great improvement in the treatment of OOV probably because the participants were explicitly aiming at domain adaptation. As a result, the quality of the cross-domain segmentation did not drop as much as expected. Closed track best word R_{oov} reached 0.87 on the finance domain in simplified characters but was only of 0.79 on the medicine domain (both simplified and traditional characters). F-scores reached 0.96 and no difference was observed between domains for which additional unsegmented data was provided and those for which no additional data was available. The invoked reason is the size of the data that was too small for unsupervised or semi-supervised algorithm to be of some help. There was still a significant drop in F-score as best F_w on 3 of the 8 closed tracks fall below 0.95 reaching only 0.94 for the medicine domain in simplified characters.

The fifth bakeoff was still very promising for domain adaptation but it may be argued that the domains involved were relatively close from one another (except for the medical corpus which lead to the lowest scores). The last segmentation bakeoff (Duan et al., 2012) demonstrates that although important progress as been made, the task cannot be considered solved in all circumstances. The 2012 bakeoff focused on microblogging data, in which the register is very different from what is usually found in manually segmented training corpora. The chosen guideline to segment the microblog corpus was the same as a larger available training corpus (of journalistic style) but the segmented data from microblog that was provided for training was very small. Most systems demonstrated a drop in performance compared to the previous bakeoff as only one system was close to 0.95 in F_w and 7

out of 17 participants fall below 0.88.

6.4. Criticism Regarding CWS Bakeoffs

Although a fair evaluation of the CWS systems was painfully needed at the end of the 20$^{\text{th}}$ century and holding bakeoffs was a great improvement in the methodology, a side-effect of the availability of corpora for evaluation is a tendency to blindy consider corpora as gold standards. Before that, discussing what kind of unit is being segmented was common in the NLP literature and some cases were explicitly considered fuzzy, see (Sproat and Shih, 1990) for an example of such work. These discussions were closely related to the linguistic issue of wordhood we addressed in chapter 3. With the availability of segmented corpora, one can simply aim at taking the same segmentation decisions as in the corpus, whatever they are. This implicitly creates a bias in favour of supervised machine learning algorithm and creates a gap between CWS in NLP and the linguistic questions. The linguist's task becomes to define segmentation guidelines of the corpora on which NLP practitioners can train various machine learning algorithm, with very few feedback to linguistic theories. All fuzzy cases described in the previous chapters are dealt with in a boolean way by the annotators while trying to be as consistent as possible. This is very similar to adding spaces to the orthography, except that only the annotators are aware of it.

The establishment of CWS as an NLP task and the organization of segmentation bakeoff allowed for a more rigorous evaluation, demonstrated the need for a detailed account of actual data that revealed some unnoticed issues. But it also established various corpora annotated with diverging guidelines as *gold standards* even though they cannot be considered as such because of the remaining linguistic issues. The segmentation of those corpora should continue to be questioned. Their availability also obfuscated the relevance of some older methods in favour of supervised machine learning algorithms. However when trying to extend the progress to new domains or registers, older rule-based or lexicon-based method are still proven useful. This is an important fact as the annotation of a new corpus for any new situation would be an incredibly expensive methodology. On top of that, as the genericity and extendability of supervised approaches is still questionable, the influence of the imperfections of CWS systems on further processing and on quantitative linguistic studies are difficult to predict.

Dong et al. (2010) even go further as they request a "radical change" in Chinese NLP, claiming that the CWS task is ill-defined and that seeking for a word-level tokenization to reach a situation close to the English NLP is irrelevant for Chinese. They underline the fact that tokenization for English does not deal with MWE and thus advocate for a character-based processing.

Our position differs from theirs. We stressed the importance of a proper treatment of complex units such as MWE in the previous chapters and we do not consider that the Chinese characters create a totally different situation from the linguistic point of view, Chinese characters may however affect the etymological intuitions of the speaker (recall the critics DeFrancis (1984), discussed in chapter 2).

Nevertheless we do agree with Dong et al.'s (2010) criticisms as part of the decisions taken to annotate the Chinese corpora resemble to *orthographic* word boundaries in Latin script as they are arbitrary, explicitly learnt and of little linguistic relevance. The fact that NLP practitioners tend to take the corpora for granted without questioning its binary aspect (or without providing qualitative analysis of the output of their systems) make many experiments in this direction irrelevant for linguistics.

As we discussed in the previous chapter, word-level tokenization for French and English was proven to be suboptimal for deeper NLP processing. Trying to reach the same word-level of segmentation for a script without word boundary punctuation marks is indeed questionable, not only for linguistic studies but also from an applicative viewpoint. Nevertheless, as we discussed previously, it is still far better than a simple character based segmentation without further grouping procedure. Although we have acknowledged the imperfection of manually segmented corpora for our task, they do provide a good basis for the evaluation and comparison of segmentation systems.

6.5. Available Corpora and Corpora Used

In this work, we use manually segmented corpora that were made freely available for the "Second International Backoff in Chinese Word Segmentation" (Emerson, 2005). They were contributed by four different research institutions which had defined four different segmentation guidelines. They are provided with an official split between training and testing data for a fair comparison of the results. Global statistics are provided in Table 6.1.

The MSR Corpus in a sample of Microsoft Research's in-house corpus. Very little information about its constitution has been provided for the Bakeoff, but more details can be found in (Gao et al., 2005). This is a balanced corpus of 40M characters in simplified Chinese containing various domains and styles. However no details were given about how it was sampled to provide training and test data for the Bakeoff.

The PKU Corpus is also written in simplified Chinese. It is part of a set of language data annotation started in 1992 at Peking University Computational Linguistics Lab (北大计算语言学研究所, CCL). The data provided for the segmentation bakeoff is specialized in style as it contains exclusively materials from the *People's Daily* (人民日報) from 1998 and follows a revised version of PKU's segmentation and tagging standard. The original corpus has also been manually tagged with parts-of-speech, but this information was removed for the bakeoff. The annotations rely on a lexicon which was also produced by CCL, the "Modern Chinese Dictionary with Grammatical Informations" (《现代汉语语法信息词典》).

The CITYU Corpus is extracted from the LIVAC corpus (T'sou et al., 1997). This corpus is gathered and processed by a team from the City University of Hong-Kong. The goal is to account for the different varieties of Mandarin Chinese from different communities. Since 1995, the team continuously sample Chinese newspapers and electronic media of Hong Kong, Taiwan, Beijing, Shanghai, Macau and Singapore. A certain degree of heterogeneity is thus to be expected in this corpus. The sample provided for the bakeoff is written in traditional Chinese.

The AS Corpus is extracted from the Academia Sinica Balanced Corpus, a corpus collected and annotated by Academia Sinica's Chinese Knowledge Processing Group (CKIP). The original corpus is also POS tagged. It is written in traditional Chinese. It follows the guidelines defined in (Chen et al., 1996) and relies on the CKIP Lexicon.

6.6. A Contrastive Overview of Segmentation Guidelines

Segmentation guidelines for all four corpora are available on the bakeoff website. We shall now provide a short overview of these guidelines.

	words		characters	
Corpus	token	types	token	types
Academia Sinica (AS)	5449698	141340	8368050	6117
City University of Hong Kong (CITYU)	1455629	69085	2403355	4923
Peking University (PKU)	1109947	55303	1826448	4698
Microsoft Research (MSR)	2368391	88119	4050469	5167

Table 6.1.: **Size of used corpora**

6.6.1. MSR Guidelines

The guidelines for the MSR corpus are the shortest. They only distinguish between lexical word, factoids and named entities. Factoids include expressions such as dates, numbers, email address or phone numbers. Named entites are names of persons, locations or organizations. With respect with the previous chapter, location and organization names are often MWE (mostly Fixed Expressions). Examples given in the guidelines include 圣海伦 岛 公园 *Saint Helen's Island park* or 毕加索 博物馆 *Picasso museum* (here the segmentation is modified to match the translation, MSR guidelines treat such MWE as a single unit). MSR's "lexical word" potentially cover all the units discussed in the previous chapters without further precisions.

shènghǎilún dǎogōngyuán Bìjiāsuǒ bówùguǎn

Guidelines for other corpora were more precise but only available in Chinese.

6.6.2. AS Guidelines

For the AS corpus, Chen et al. (1996) and Huang et al. (1996) provide the description of the methodology followed to build the Sinica Corpus from which the AS Corpus of the bakeoff is sampled. AS guidelines are composed of two general principles and a set of auxiliary guidelines to use for unclear cases. They define a *Segmentation Unit* as the "smallest string of character(s) that has both an independent meaning and a fixed grammatical category". The principles state that:

1. A string whose meaning cannot be derived by the sum of its components should be treated as a segmentation unit.

2. A string whose structural composition is not determined by the grammatical category other than the one predicted by its structural composition should be treated as a segmentation unit.

These correspond to the *semantic composition* and *exocentric structure* criteria for which we saw in chapter 3 they may yield unclear results. The definition of the segmentation unit itself is problematic given the mismatch of minimal units described in (Kahane, 2008) and discussed in chapter 3. The auxiliary guidelines are less categorical, they are heuristics to make consistent decisions about unclear cases. These are formulated as follow:

1. Bound morpheme should be attached to neighbouring words to form a segmentation unit when possible.
2. A string of characters that has a high frequency in the language or high cooccurence frequency among the components should be treated as a segmentation unit when possible.
3. Strings separated by overt segmentation markers should be segmented.
4. Strings with complex internal structure should be segmented when possible.

These guidelines are meant to be provided with a lexicon that list words and bound affixes. The "when possible" of the first guideline hides difficult cases related to phrasal affixes and clitics. The second guideline may be difficult to apply to MWE. The third one is clear, except for the semi-coma (、) that allow coordination of bound morpheme as in 火、氣車站 (huǒ,qìchēzhàn) *train and bus station* (see section 3.2.1 for a complete discussion).

The authors of the corpus were well aware of these difficulties and had both linguistic and practical NLP applications in mind so they propose three levels of segmentation standard by increasing difficulty order. They give the following formulation:

Faithful All segmentation units listed in the standard lexicon should be successfully segmented

Truthful All segmentation units identified at the Faithful level as well as segmentation units derivable by morphological rules should be successfully segmented.

Graceful All linguistic words are successfully identified as segmentation units.

However, the current target in CWS evaluation in open tracks is somewhere in between Truthful and Graceful. Some OOV word may be correctly segmented

even if they are not derivable from some official lexicon using a morphological rule. Closed tracks do not allow for a full external lexicon, in that case even the *Faithful* level may not be reached. On the other hand, we are not aware of any experiment that would prove a segmentation system to be "perfectly Truthful". The linguistic word at *Graceful* level remains undefined.

The documents provided for the 2005 Bakeoff[1] include precise guidelines along with numerous examples. It also include a list of affixes and a table summarizing the decisions for a list of difficult cases. The auxiliary guidelines slightly differ from (Huang et al., 1996):

1. If there is an obvious separation mark, segmentation is required. (This includes insertions[2] and punctuation.)
2. Bound morpheme should be glued with their left or right context as much as possible. (with the exception of clitics 的 de 地 de 之 zhi .)
3. Sequences with high frequency of occurrence or co-occurrence should be considered as a single unit.
4. Verbs made of two characters with a modifier-modified internal structure should be considered as a single unit. (These are often frozen expressions)
5. Noun made of two characters followed by a noun made of one character with a modifier-modified structure should be considered as a single unit.
6. Expressions with a complex internal structure should be segmented.

6.6.3. CITYU Guidelines

The guidelines provided for the CITYU corpus first distinguish specific treatments of proper names, words for numbers, date and abbreviations. Then they give indications classified by the length of an expression.

Names of person and country are considered as one unit (even for country names which translate into MWE such as 中華 zhōnghuá 人民 rénmín 共和國 gònghéguó [3]'People's Republic of China'

[1] All the other documents related to the segmentation guidelines are in MSC, translations in the remaining of this section are ours.
[2] The potential clitics 得 de and 不 bu presented in chapter 3 are considered insertions.
[3] We segment to match the translation, the corpus does not.

(lit. *Chinese-People-Republic*). A list of expressions for *street*, *county*, *river*, ...is provided, monosyllabic expressions are considered as suffix (they form one unit with the name of the street or the city) but longer expressions form a unit by themselves, given example include 香港 • 地區 (xiānggǎng • dìqū) 'Hong-Kong area' 廣西 • 自治區 (guǎngxī • zìzhìqū) 'autonomous region of Guangxi'.

Locative particules[4] (that are analysed as phrasal affixes by Liu (1998)) are glued to monosyllabic words but are standalone segmentation units when they are associated with longer expressions, contrasting 家裏 (jiālǐ) 'at home' and 學校 • 裏 (xuéxiào • lǐ) 'at school'.

The general guidelines includes syntactic regularity and semantic compositionality considerations, but also a factor of frequency. For example 牛肉 (niúròu) *beef* is one word but 鹿 • 肉 (lù • ròu) *venison* is segmented (牛 (niú) means 'bovine', 鹿 (lù) 'deer' and 肉 (ròu) meat).

A list of productive suffixes (bound morphemes) is provided but the division of the guidelines according to the length of expressions is problematic. It states that bisyllabic words must be glued with the suffix as in 在野黨 (zàiyědǎng) *opposition-party* but there may be cases where "a bound morpheme stands as a free form"[5], as in 本 • 黨 (běn • dǎng) 'This party'. On top of that, if it is attached to a phrase it may be segmented into a standalone unit, as in 民主 • 進步 • 黨 (mínzhǔ • jìnbù • dǎng) 'Democratic Progressive Party'.

It is also important to note that the list of productive affixes does not perfectly match the one provided for the AS Corpus.

CITYU guidelines also require to glue together idioms and formula if they are syntactically frozen (originating in classical Chinese syntax or modern syntax) or if they are common expression formed regularly but with specialized meaning, for example in politic language as 一國兩制 (yīguóliǎngzhì) *"one country, two systems" policy (that describes the relation between China and Hong-Kong).*

The CITYU Corpus originally aims at accounting for variation of MSC among different communities, code switching and borrowings from other sinitic languages is considered and shall not be segmented. A distinction could have been made as if gluing together borrowed expressions sounds reasonable, it could be a better choice to perform proper segmentation of the other language in case of code switching. But practically speaking, the amount of data in other languages is probably not fitted for this task.

[4]方位詞
[5]一個黏著語素是相當自由的

6.6.4. PKU Guidelines

The PKU segmentation standard is mostly based on the lexicon and a set of instructions for specific cases. It states that "a Segmentation Unit have defined meaning and syntactic function".

The length of the unit is also explicitely mentioned in the guidelines: *If we consider the number of characters, two-characters units are common, three characters unit are more unlikely and the longer strings are generally segmented into smaller units except for idiomatic expressions.*

Most of the segmentation decisions will follow the lexicon. As it contained 70,000 entries in 1999, it was expected to cover the vast majority of segmentation units in the corpora to be processed. The guidelines state that: *any expression that corresponds to an entry in the lexicon should be considered as a segmentation unit (including: words, phrases, idiomatic expressions, abbreviations…*

After stating general principles, the guidelines give instructions for a list of specific cases. We comment only on a selection of items.

1. First and last names of Chinese names should be splitted
2. Transliterated foreign names form a single unit (even if it include a middle dot)
3. Name of famous authors, or *nom de plume* for which it is difficult to distinguish first and last names form a single segmentation unit.
4. Country names always form a single segmentation unit
5. The single characters indicating a kind of location form a single unit with the name of the place. The list is 省市县区乡镇村旗州都府道 江河山洋海岛峰湖街路道巷里町庄村弄堡.
6. Longer locative expressions are standalone segmentation units.[6]
7. Organisations names that are in the lexicon form a single unit.
8. Organisations names that are not in the lexicon and form a decomposable expression (especially when they contain a place or persone name) should be segmented.

[6] bracketed in the tagged corpus

9. Potential forms of verbs (with 得 de or 不 bu) are segmented in three units as in 走 zǒu • 得 de • 到 dào 'being able to arrive by walking' but not when the potential marker is mandatory and removing it would result in an invalid form or a different meaning, as in 说得过去 shuōdeguòqù 'being justifiable' / * 说过去 shuōguòqù.[7]

10. compounds made of two verbs form a single unit if it is recorded in the lexicon or if it can't be decomposed into two autonomous verbs.

11. locative phrasal affixes form a separated segmentation unit, except when they form a word with a non-free morpheme.

The full annotation schema allows for bracketing of complex expression which may lead to the same kind of discrepencies as the CITYU guidelines. The examples given include the following analyses:

国防部/nt *ministry of defence*, where 部 bù is a suffix for ministries and [信息/n 产业/n 部/n]nt*ministry of information industries* which in the bakeoff PKU corpus is turned into 信息•产业•部 where 部 bù is a standalone unit. (In the tagged corpus, */n* marks a nominal expression and */nt* an nominal expression with locative meaning.)

The same bracketing applies for place names but does not seem to create discrepancies. Single character location types will not be segmented as in 北京市 běijīngshì 'Peking municipality', but more complex expressions will: [香港/ns 特别/a 行政区/n]ns which becomes simply 香港•特别•行政区 'Hong-Kong Special Administrative Region'.

6.6.5. Remarks

These different guidelines and heuristics yield mostly consistent segmentationsbut disagree in many cases. Xia (2000) claims that the disagreements between various guidelines are a matter of a limited set of conventions and that it should be straightforward to design a rule-based system to convert a corpora to another guidelines. This is true for the aspect of segmentation that are strictly arbitrary such as segmentation of factoids (date, proper names…). It is less likely for the decisions that rely on judgements in productivity of affixes or degree of compositionality of meaning.

In any case, the binary segmentation used in the bakeoffs fails to account for syntactic and semantic discrepancies and do not exhibit a proper treatment of phrasal

[7] The very common expression 對不起 duìbuqǐ 'sorry' is of this kind. It is a frozen formula whose non-potential counterpart is non longer in use.

affixes. They provide a good common ground for a unified evaluation and allowed important advances in NLP, but they are unable to account for the aforementioned "Graceful" level of Huang et al. (1996).

These heuristics allow a consistent segmentation of corpora, but a segmentation that may conflict with the criteria presented in chapter 3. This is mostly due to the fact that in CWS, the clear-cut segmentation is in charge of the identification of both semantic and syntactic units in a unified way even though they could be conflicting. The important role played by frequency and length is also likely to result in structural inconsistencies.

Chapter 7

Overview of Word Segmentation Systems

7.1. Systems Relying on External Resources

7.1.1. Lexicon-based Algorithms

Some of the earliest works in CWS fall into the category of lexicon-based algorithms. The only required resource is a lexicon which may be feature-rich or a simple list of words. An intuitive way to model this family of algorithm is to consider the boundary positions between characters and to define a directed acyclic graph (DAG) with the boundary positions as vertices and the lexicon entries as edges. The simplest graph containing only the correct solution for the previous example is illustrated on Figure 7.1. It is build as if only the expected words were part of the lexicon. Most of the time, ambiguities will arise with more complete lexicon and a realistic example of such a graph is presented Figure 7.2.

Once modelled as a DAG, performing CWS can be reformulated as finding the correct path in the DAG. A straightforward yet quite efficient algorithm is called *Maximum Matching*. It is a greedy algorithm that starts from the right (forward matching) or from the right (backward matching) and from any position reached in the sequence to be segmented will select the longuest known word in the lexicon to reach the next position. A specific heuristic needs to be used to deal with Out-Of-Vocabulary words to ensure at least one path from the first to the last positions.

	AS	CITYU	MSR	PKU
Baseline	0.882	0.833	0.933	0.869
Topline	0.982	0.989	0.991	0.987

Table 7.1.: **Baselines and Toplines on Bakeoff 2 corpora with Maximum Matching algorithm**

Figure 7.1.: **Simplest DAG (expected segmentation)**

The edges can also be weighted or labelled to add lexical or stochastic information that can be used to select the segmentation path. An example of such algorithm mixing lexical and stochastic data to perform segmentation is presented in (Sproat et al., 1996) where the authors model the lexicon itself as a weighted finite state transducer that is combined with the input sequence to build an automaton.

These approaches now serve to define *baseline* and *topline* for closed evaluation of CWS. The baseline is a Maximum Matching algorithm using the lexicon extracted from the training corpus and the topline uses the same algorithm but with a lexicon extracted from the test corpus. One the bakeoff 2 corpora, this yield the results presented in Table 7.1.

7.1.2. Supervised Machine Learning

A large variety of segmentation systems based on supervised machine learning has been proposed. Most of the Machine Learning families have been tried since the first segmentation bakeoff of "standard" corpora. When given consistent training and testing data, F-score over 0.97 can be reached

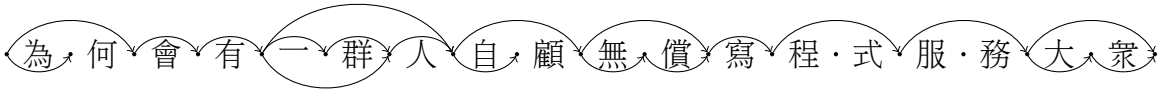


Figure 7.2.: **A more realistic DAG**

The goal of supervised learning is to enable the computer to make decisions similar to a large set of examples. In this context, an example is a pair made of an observation and a decision that was taken, typically an annotation made manually.

The good results prove a relatively high degree of consistency in the manual segmentation of the corpora. Nevertheless, there are two major issues that may make such algorithm unfitted for linguistic inquiry:

- Supervised systems are bound to a specific training data, following specific annotation guidelines.
- Their errors are unpredictable, especially when we want to proceed texts that differ a lot from the training corpus.

Annotating a training corpus is a labour intensive task. It would be even more costly if we were aiming at a high degree of linguistic felicity. Therefore they are not a good choice to analyse data that is not similar enough to available training data and if we need more than binary segmentation. Note that there is no easy way to define what "similar enough" means and the only way to know if a segmentation system perform well on a specific kind of data is to manually annotate a sample for evaluation.

The fact that supervised learning relies on manually annotated examples contradicts our goal to agnostically explore the data to find an emerging segmentation. For all the unclear cases on which linguists may disagree, the annotators followed the heuristics presented in the previous section to make a consistent choice. The high f-scores reached by some systems tend to show that the choices were indeed made with a high level of consistency, otherwise the machine learning algorithms wouldn't be able to follow the guidelines systematically. Nevertheless, we consider these choices as arbitrary as *orthography* can be in terms of spacing in English or French. This corresponds to an *orthography* that only the annotators were trained to follow. As such, its linguistic relevance is questionable. The good or wrong answers by CWS systems on the consistent yet arbitrary choices made by the annotators are of less importance. Unfortunately, it is difficult to estimate the proportion of arbitrary splits to linguistically motivated splits in a corpus.

In section 7.2.6 we use a supervised segmentation system based on a perceptron and available as off-the-shelf as an open source software to estimate the consistency

of the differents guidelines (Zhang and Clark, 2007, 2011).

7.1.3. Adaptive Supervised Systems

To account for the variety of segmentation standards or to help with domain adaptation, hybrids methods that combine machine learning and rule-based algorithms have been proposed (Wu, 2003; Gao et al., 2005).

Wu (2003) explicitly gives up linguistic correctness: "we do not have to wait for linguists to reach a consensus before we do segmentation in NLP". He aims at addressing the need of different granularities of segmentation units for different NLP applications. In order to do so, he remarks that most of the disagreements between the different guidelines lie in morphologically derived words (MDWs) and factoids. He proposes a system that provides both segmentation and internal structures of the MDWs and factoids. Internal structures are given by a lexicon with morphological information and a set of rules to derive new MDWs. Internal structures are represented as trees and are used to alter the segmentation with a set of about 50 boolean parameters (for example: whether or not first name and given names shall be segmented). He achieves high results (f-score $> .96$) on different corpora from the first bakeoff by manually changing these parameters. Gao et al. (2005) improve on this system by using a Linear Mixture Model (LMM) to obtain an initial segmentation with internal structure for MDWs, factoids and NEs. They use a combination of Microsoft's lexicons and an in-house corpus (which follows MSR guidelines but is ten times larger) to train the LMM. Transformation Based Learning (Brill, 1995) is then used to automatically adapt the output of the system. The basic system obtains f-scores from 0.820 to 0.839 (depending on the corpus), with TBL adaptation they reach scores from 0.954 to 0.958.

Despite the high results achieved on each standard, the choice of the proper annotation scheme for a given NLP task is still delicate. Proper evaluation and tuning is required for each particular task. An extreme case is studied by Sun and Lepage (2012) who demonstrate that better results can be achieved in Machine Translation between Chinese and Japanese when no segmentation is performed prior to the extraction of translation candidates. In that case, both languages do not have explicit marking of word boundaries. The phrase alignment stage of the Machine Translation system performs an implicit segmentation based on the two languages at the same time.

From a linguistic point of view, the efficient methodology followed by Gao et al. (2005) cannot compensate for the drawbacks of the binary segmentation of the training corpora. It seems to us that unsupervised learning is needed for an agnostic exploration of the language data.

7.2. Unsupervised Systems

In this section, we will review works in Word Segmentation that do not rely on manually annotated training data nor external lexicon. These works require only raw data and try to induce its segmentation from the distributions of forms observed in the corpus. They constitute the state-of-the-art on which our own work is based and which it aims at outperforming.

Although these methods usually perform worse than systems based on manually prepared resources, we believe that they are more likely to provide linguistically valuable insights about wordhood. They also have the benefit to limit the need for manual work on the data. This make them less expensive and more domain independent.

The first work in this area was done by Sproat and Shih (1990). As it is a pioneering work relying on a simple measure of cohesion, we think it is worth to describe it here.

Beside cohesion measure, works in unsupervised segmentation typically rely on some separation measure that are often related to Harris (1955, 1967) hypothesis, to a combination of cohesion and separation measures or to some estimation of the probability of the observed sequence of characters given a segmentation. Such probability is generally estimated using *Bayesian inference*. Another framework for unsupervised word segmentation is derived from Information Theory paradigm of *Minimum Description Length*. The latter two being mathematically related.

7.2.1. Sproat and Shih, 1990

The first work of this kind we are aware of is the one by Sproat and Shih (1990). At that time no training corpus was available and computer-readable lists of MSC words were too small (they mentioned a list of 6000 words). Following Suen (1986), they claim that 69% of Chinese texts consist of one character words and that only 1% consists of words of three characters or more. For that reason, they focus only on

grouping together words made of two characters. The statistics we can draw from large segmented corpora are quite different: only 22% of the PKU training corpus consists of one character words and 62% consist of two characters word. As a result, 15% of the corpus have no chance to be correctly segmented by their system, not 1%. Nevertheless this is an important pioneer work.

To group a pair of characters as a word, they define a metric inspired by the Mutual Information (MI), a metric from Information Theory that use probabilities to model how strongly related two event are.

$$MI(a;b) = \log_2 \frac{P(a,b)}{P(a)P(b)}$$

If the probability $P(a,b)$ to observe a and b together is significantly higher than the probability $P(a)P(b)$ to observe a and b together under the assumption that the two events are independent, there is a good reason to believe that a and b are related.

They estimate the probabilities based on the numbers of occurrences of a, b and ab in the corpus of length N to define their association measure:

$$A(ab) = \log_2 \frac{\frac{\#ab}{N}}{\frac{\#a}{N}\frac{\#b}{N}} = \log_2(N) + \log_2 \frac{\#ab}{\#a\#b}$$

where $\#x$ stands for "number of occurrences of x in the corpus".

Unlike MI, their association measure is sensitive to the order of occurrence, as observing the string ab in the corpus is not the same as observing ba.

They compute the association of all bigrams in the sequence to be segmented. They provide the following example, when segmenting the string 我弟弟現在要坐火車回家, they obtain the following statistics:

a	*b*	*A*(*ab*)
我	弟	0.00
弟	弟	10.44
弟	現	0.00
現	在	4.23
在	要	-2.73
要	坐	0.00
坐	火	0.00
火	車	7.31
車	回	2.06
回	家	4.69

They group characters two by two by decreasing order of association, only if the grouping is not blocked by a previous grouping with higher association and until they reach a threshold of 2.5 (selected manually). In the example, the grouping order will thus be: 弟弟 (dìdì), 火車 (huǒchē), 回家 (huíjiā) and 現在 (xiànzài). This leads to the correct segmentation:

(1) 我 弟弟 現在 要 坐 火車 回家
wǒ *dìdì* *xiànzài* *yào* *zuò* *huǒchē* *huíjiā*
I younger brother now will sit train return home
'my younger brother is about to take the train back home'

They tried this algorithm with a corpus of 2.6M characters and provide results in terms of boundary precision and recall, but counting only clear cases of split and clear cases of words of two characters. They report a precision of 0.90 and a recall of 0.94, but those figures are not comparable with recent bakeoff as they focus only on the relevant subset of boundaries and do not penalize ambiguous cases.

7.2.2. Bayesian Inference and Psycho-linguistic Researchs

An important part of the research in unsupervised word segmentation is in fact unrelated to Chinese. It focuses on the role of various cues for word segmentation in language acquisition by humans. Infants are able to learn words from continuous speech. The main question is whether this learning process can be made solely by exposition to a language or if it require some innate knowledge. Saffran et al. (1996) for example explore to what extent segmentation can be inferred from distributional cues, especially transitional probabilities. They do so with psycho-linguistic experimental methodology.

This research on language acquisition raises the question of acquisition modelling. Being able to model a theory of acquisition is important to prove the learnability of the data (and reject a nativist hypothesis). It may also allow to compare the relative importance of different cues. Brent (1999) proposes a model capable of integrating phonology, word-order, and word frequency cues in a modular fashion. He underlines the differences between acquisition models and systems engineered for the task of word segmentation in NLP, stating that acquisition models:

- must start out without any knowledge specific to a particular language;
- must learn in a completely unsupervised fashion;
- must learn and segment incrementally, as a first approximation, this means that the segmentation of each utterance must be finalized before the next utterance is read in;
- and that the cognitive modeling goal dictates the kind of corpus [...] —phonemic transcripts of spontaneous speech by mothers to their young children (often called "Child Directed Speech").

Concerning the first point, one may argue that the fact that we are using corpora written with Chinese characters make our work language-specific. This is true to some extent but preliminary experiments show that it can be extended to a variety of languages.

The second point applies for our goal. We refuse to use any external resource such as lexicon or manually annotated corpora.

The third constraint is not needed for our goal. We allow ourselves multiple passes on the corpus. We aim at finding a way to analyse the corpus, not to model the acquisition or reading process.

The fourth point is the biggest difference. We work on textual data as it is commonly written in different contexts. Note however that using phonemic transcripts is already a kind of supervision as phonemic transcription from speech signal is a non-trivial linguistic processing. An important side effect of this difference is the size of the data we are dealing with. The large majority of work on unsupervised segmentation in acquisition modeling is done on the Bernstein-Ratner corpus (Bernstein-Ratner, 1987), hereafter BR corpus, which consists in only 9,790 utterances, 33,387 words and 95,809 phonemes. The small size of the corpus allowed the researchers involved in this track to develop more complex algorithms that are

practically unusable on the dataset for CWS. In fact, the possibility for a human to learn a language with such a small input is questionable. The fact that the BR corpus is made of Child Directed Speech also limits the vocabulary size. It contains only 1,380 different words. Short utterances and repetitions are also very common, as a result the corpus contains only 5,900 distinct utterances.

A large body of the research in this direction is done under a Bayesian learning framework. A review and important advances can be found in (Goldwater, 2006; Goldwater et al., 2009).

The general idea is to use Bayes' rule to define the probability of some hypothesis h (a model or a linguistic hypothesis responsible the generation of the corpus) given the observed data d:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

As the observed data is fixed by the corpus, $P(d)$ is a constant. Therefore

$$P(h|d) \propto P(d|h)P(h)$$

Finding the best hypothesis is thus equivalent to find the h that maximises $P(d|h)P(h)$. That is the product of the probability to observe the data under the hypothesis h (generally called the *likelihood*) by the probability of the hypothesis h (called the *prior*).

$P(d|h)$, the probability that the hypothesised model generates the observed corpus is typically easier to compute than $P(h|d)$ as we are free to define how the model generates the data.

If we decide to use a constant prior, $P(h) = k$, the problem becomes equivalent to a *maximum likelihood*, that can be found through Expectation Maximisation (EM) algorithm. But the equiprobability of all the hypothesis in the search space is very unlikely. The *maximum likelihood* estimation is likely to result in overfitting data, that is to favour complex and unlikely hypothesis that perfectly fit the observed data, but unable to generalize over unseen data.

Goldwater (2006); Goldwater et al. (2009) proposes a two-stage bayesian model framework. To account for observed data, the model includes a *generator* responsible for the creation of lexical items and an *adaptor* responsible for the distribution of the lexical items in the corpus enforcing a power-law distribution on word frequencies.

She suggests to use a Dirichlet process or a Pitman-Yor process to obtain sensible priors. This model is then used to perform segmentation of the BR corpus. Two models are built based on the same framework, the first one is based on a Dirichlet Process (DP) and assumes that word probabilities are independent of context (it corresponds to an unigram language model) and the second one is based on a Hierarchical Dirichlet Process (HDP) which takes the context into account with a bigram model. DP results in undersegmentation but HDP outperforms all previous works on the BR corpus.

Johnson et al. (2007) and Johnson and Goldwater (2009) generalise DP and HDP with Adaptor Grammars (AG) that provide a way to associate probabilities to a subset of the rewriting rules of a Context Free Grammar (CFG) using similar Bayesian inference. They show that one can (manually) write small grammars with which AG is equivalent to DP and HDP and yields results consistent with those reported by Goldwater et al. (2009).

The main drawback of this approach is that it is computationally intensive. The BR corpus is small compared to the corpora we may want to process (including CWS bakeoff corpora) that are intractable for DP, HDP or AG. Mochihashi et al. (2009) propose a faster algorithm based on a Nested Pitman-Yor Language Model (NPYLM) that allows it to process CWS bakeoff data and achieve some of the best results in unsupervised CWS. They report F_w of 0.807 and 0.817 on MSR and CityU corpora respectively, but results on other corpora are not provided. Unlike AG, the source code for NPYLM has not been released. No precise figures are provided but it seems that although the complexity of NPYLM is reduced, it is still computationally intensive compared with other unsupervised methods designed to segment large corpora. Published figures are not easily comparable, nor is the precise complexity of the algorithms. But on the BR corpus HDP is reported to run in about 12 hours and NPYLM in 17 minutes. The systems we propose in the second part of this thesis can process the BR corpus in a few seconds (see chapter 12).

A side effect of this drawback is that being limited to small dataset make it very hard to observe the influence of the amount of data used to train the system.

Pearl et al. (2010) underline the fact that Bayesian learning provides a *rational* model that does not take into account the limitation of human cognition and can find a solution that is optimal given the data using computational procedures that human cannot use. Humans may thus reach a sub-optimal solution. They propose to

constrain the models with human-like limitations such as working memory size. They show that such constrains may prevent some of the clues found in the literature to be usable and that constrained learners can reach higher scores than ideal learners. We think that this kind of result may vary dramatically depending on the size of the processed data. Due to data sparsity it seems likely to us (however hard to prove) that a small data set will present an optimal solution more distant to the actual human-like solution than a larger dataset.[1] Experiments have been made to compare human performance and statistical models by Frank et al. (2010) but only on artificial languages.

Although we are looking for a segmentation that is related to human cognition, we are not trying to simulate the learning process and do not have to impose such limitations on our system.

Johnson and Demuth (2010) experimented AG on MSC data but focused on phonemic transcripts of the CHILDES corpus to be of the same kind as the BR corpus. Results are therefore not comparable with our work.

Unfortunately, this research direction suffers from the same drawback as CWS. The widely used BR corpus is segmented according to the orthographic segmentation of written English. Although psycho-linguistic researches focusing on language acquisition and mental lexicon tend to show that the mental lexicon units and orthographic units do not match (Sosa and MacFarlane, 2002), orthographic segmentation is considered as a gold standard and used for evaluation of the acquisition models without further questioning. It seems to us that the part of the gold segmentation that in fact corresponds to explicit learning through orthography should not be overlooked.

When dealing with languages without orthographic word boundary, a similar methodology is used, like by Johnson and Demuth (2010) on MSC or Fourtassi et al. (2013) on Japanese where the authors don't question the nature of the units they inherit from the manual segmentation of the corpus.

In this case, there is no explicit learning of an orthographic segmentation but what would be a sensible *topline* given the inter-speakers agreement rate should be discussed (we will give more details on this point in section 7.2.6). However, we have not found such a discussion in the literature.

This led Fourtassi et al. (2013) to the questionable conclusion that concerning

[1] This is consistent with our findings when using Minimum Description Length (MDL), and MDL can be seen as a kind of Bayesian inference (see section 7.2.4 and chapter 10 for details).

word segmentation, *English was intrinsically less ambiguous than Japanese* (sic). A more plausible and appealing interpretation of the results they provide is that the various *Levels* of the CFG they use with AG may fit different typological properties and/or different kinds of units. Theses differences could only be captured if the evaluation procedure was accounting for *weakly autonomous* and *indissociables* forms. This is not the case with an evaluation based on a binary segmentation that tries to mimic orthographic words in English.

Although the Bayesian approach yields the best results in many tasks, it does not exactly fit our needs, even if the issues concerning training time are ignored. All the models we mentioned in this section require manual settings of some parameters to define the prior distributions of the hypothesis. The output of these systems are not trivially translated into an approximation of the syntactic freedom. We want to avoid parameter settings as much as possible and propose a method to empirically measure some kind of autonomy. A solution could be to carefully design a CFG capable of capturing the various kinds of units discussed in chapter 3 but it would have to be hand-crafted. In this work, we target a systematic measure of the data which is as objective as we can, fitting an handwritten grammar conflicts with this goal.

7.2.3. Harris

In his article "*From Phoneme to Morpheme*", Harris (1955) makes the hypothesis that the morpheme boundaries are related to the variety of possible successors following a sequence of phonemes.

More precisely, he postulates that if we take prefixes of increasing length of a given utterance, the number of possible following phonemes will decrease, except when we reach a morpheme boundary after which a larger variety of phonemes may eventually occur. The related idea is that the longer the prefix, the more predictable the next phoneme for a native speaker of the language, except when we reach a morpheme boundary where the predictability suddenly decreases (or the number of possible following phonemes increases).

He designed a procedure to perform the segmentation of a sequence of phonemes systematically based on the successor variety after each phoneme. The basic rule is to segment whenever the number of possible successors is at a local maximum. For example, the successor variety after each phoneme for the utterance "*he's quicker*" (phonemized as */hiyzkwikər/*) is given in Figure 7.3 where the red lines indicate the

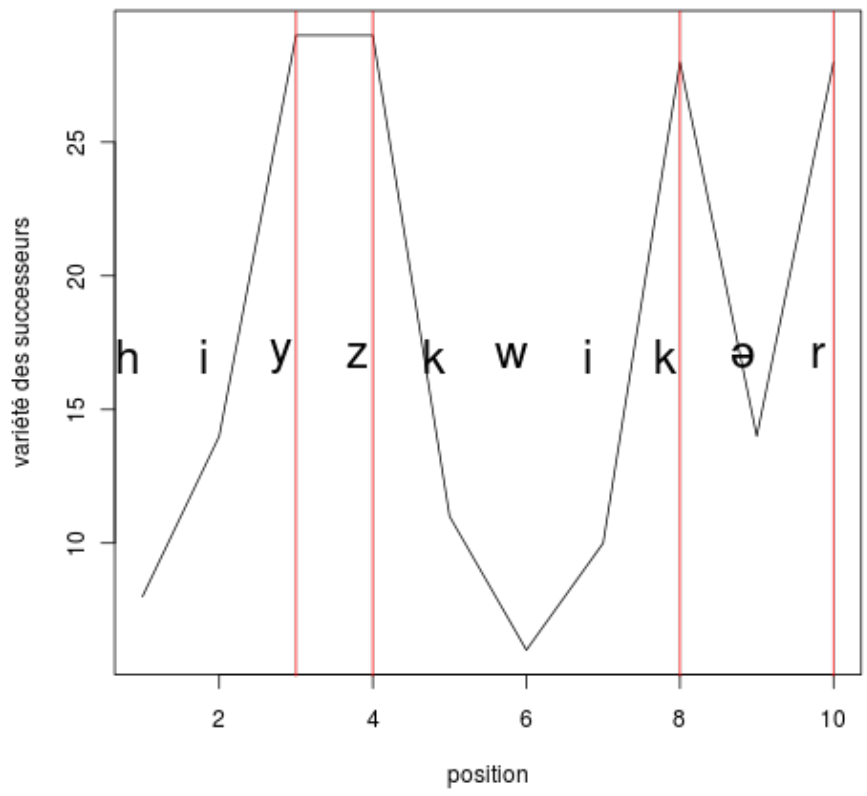


Figure 7.3.: **Segmentation of "*He's quicker*", following Harris (1955)**

segmentation.

To obtain the values for the successor variety, Harris did not rely on a corpus, not available at this time, but on speaker elicitation, he explicitly asked native speaker how many phoneme they can think of that can follow the different prefixes.

Harris' hypothesis is at the root of many works in word segmentation, including pyscholinguistic Saffran et al. (1996) and (supervised, unsupervised or semi-supervised) NLP (Kempe, 1999; Feng et al., 2004; Zhao and Kit, 2008; Zhikov et al., 2010).

NLP works that follow Harris hypothesis fall into two categories: those who stick to the initial formulation as a discrete number of successors or predecessors of a

sequence, refereed to as *Accessors Variety* (AV) like in (Feng et al., 2004) and those who use its formulation in terms of *Branching Entropy* (BE). BE was first used by Kempe (1999) and can be considered as a continuous version of AV that takes into account not only the variety of accessors but also their respective frequencies in a corpus.

We think that the Entropic formulation of BE makes more sense for our goal because it can attenuate the noise coming from the corpus and it is a better model of predictability.

Another distinction can be made between systems that directly use AV or BE and the systems that use the variation of AV or BE between one position and the next (or the preceding) one. Using the variation (or derivative) is closer to the initial hypothesis by Harris.

A promising system was proposed in (Jin and Tanaka-Ishii, 2006; Jin, 2007) for unsupervised CWS, and is the closest to Harris' hypothesis in using the variation of BE (hereafter VBE). It inspired our own systems proposed in the next part of this thesis.

Let us consider χ the set of all possible characters and $P(x|x_n)$ the probability that the character x follows the n-gram x_n.

BE can be defined as follows

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

and VBE is then simply

$$VBE(x_n) = BE(x_n) - BE(x_{n-1})$$

The n-grams can be counted by reading the corpus from left to right (forward) to get the right VBE or from right to left (backward) to get the left VBE.

The system proposed by Jin and Tanaka-Ishii (2006) uses n-gram language models with $1 \geq n \geq 6$, in both reading directions to compute 12 values of BE and VBE at each position between two characters. Based on these values, three rules are used to decide if a given position is a boundary:

- The BE is at a local maximum
- The VBE is positive or greater than a given threshold

- The BE is higher than a given threshold

Whenever one of these three conditions is met, the location is considered as a boundary.

A comprehensive evaluation of the system is provided in (Jin, 2007). On a corpus similar to the PKU corpus used in the Bakeoffs (journalistic style sampled from the "People's Daily") it achieves a F_w of only 0.72. However, the learning curve shows a constant P_b around 0.9 and an increasing R_b that has not converged to its maximum when using the whole dataset. The error analysis is also promising. Subjectively, the reported errors seems to "make some sense" from a linguistic point of view and deserve further investigation.

The main difference between Kempe's and Jin's systems is that Kempe calculates the VBE using contexts of constant length where Jin uses contexts of increasing length. This difference is discussed more in details in section 8.2.2

The principal limitation of Jin's system for our purpose is that it acts as a binary classifier on the boundaries. We aim at defining a continuous measure on the wordforms. However, we think that the underlying hypothesis and the VBE deserve more investigations. This will be the focus of the second part of the dissertation.

Another limitation is the locality of the decision. Jin's base system does not use global information drawn from the sentence or corpus level. To compensate for this limitation, he adds a procedure based on Minimum Description Length (MDL) that allows him to reach a F_w of 0.78. More will be said about the use of MDL in CWS in the next section.

7.2.4. Minimum Description Length

The Minimum Description Length was introduced by Rissanen (1978). It can be considered as an approximation of the Kolmogorov complexity or as the formalisation of the principle of least effort (Zipf, 1949) as a compression model. We will see that it can also be seen as a specific case of Bayesian inference.

The underlying idea behind the use of MDL for Word Segmentation is the following: once a corpus is segmented, it can be recoded as a lexicon and a sequence of references to the lexicon. A good segmentation should result in a more compact representation of the data. Probability distributions of lexical items in the corpus and Shannon entropy from Information Theory are used to determine the theoretically optimal compression rate we could achieve with a given segmentation

(we just need to estimate the probabilities to obtain the theoretical value, not to actually perform the compression).

Formally, a segmented corpus will be considered as a sequence of codes referring to a lexicon, or word model, M_w (sometimes also called a *codebook*), which represents each word using a code that depends on the frequency of the word in the corpus: to get a smaller representation of the corpus, a frequent word is to be represented by a shorter code.

The description length $L(C)$ of a corpus C can then be computed as the length $L(M_w)$ of the lexicon plus the length $L(D|M_w)$ of the sequence of word codes to account for the corpus data D given a lexicon M_w:

$$L(C) = L(D, M_w) = L(D|M_w) + L(M_w).$$

The content of the lexicon itself (which can be seen as a list of wordforms D_w) can be further encoded as a sequence of characters, using a model M_c accounting for characters probability distributions in the lexicon. Again, more frequent characters will have shorter codes. As a result,

$$L(M_w) = L(D_w, M_c) = L(D_w|M_c) + L(M_c).$$

Information Theory tells us that for a given segmentation, $L(D|M)$ the optimal compression of some data D using a codebook M is given by:

$$L(D|M) = -\log_2 P(D|M).$$

which under the unigram assumption can be computed as follows.

Let us consider N the number of items in the Data sequence, $|M|$ the size of the codebook, $\#w_i^D$ the total number of occurrences of the i^{th} item, indexed by the Data sequence order and w_j^M the j^{th} item, indexed in by the codebook order. In other words, if we consider the code length of the sequence of words that form the corpus, $\#w_i^D$ will be the number of occurrences of the i^{th} word in the corpus and w_j^M will be the j^{th} word in the lexicon.

$$L(D|M) = -\log_2 \prod_{i=1}^{N} \frac{\#w_i^D}{N},$$

$$L(D|M) = -\sum_{j=1}^{|M|} \#w_j^M \log_2 \frac{\#w_j^M}{N},$$

All the required values are straightforward to count in a segmented corpus.

As shown for example by Zhikov et al. (2010), it is possible to decompose this formula to allow fast update of the DL value when we change the segmentation and avoid the total computation at each step of the minimization.[2]

MDL is often used in unsupervised segmentation systems, where it mostly plays one of the two following roles: (i) it can help selecting an optimal parameter value in an unsupervised way (Hewlett and Cohen, 2011), and (ii) it can drive the search for a more compact solution in the set of all possible segmentations.

When an unsupervised segmentation model relies on parameters, one needs a way to assign adequate values to them. In a fully unsupervised setup, we cannot make use of a manually segmented corpus to compute these values. Hewlett and Cohen (2011) address this issue by choosing the set of parameters that yields the segmentation associated with the smallest DL. In their experiments, the output corresponding to the smallest DL almost always corresponds to the best segmentation in terms of word-based f-score. Although this is a sensible hypothesis, we will show in chapter 10 that it may be proved wrong provided a sufficiently large search space (which may not be the case when MDL is solely used for the estimation of a limited set of parameters).

In the system by Zhikov et al. (2010), the initial segmentation algorithm requires to chose a threshold: for a given position in the corpus, they simply mark the position as a word boundary if the BE is greater than the threshold. The value of this threshold is unsupervisingly discovered with a bisection search algorithm that looks for the smallest DL.

When using MDL to guide the search for a good segmentation amongst all possible segmentations of a corpus, the main issue is that there is no tractable search algorithm for the whole hypothesis space. One has to rely on some heuristic procedure to generate hypotheses before checking their DL and choose the best one. This is commonly done in a iterative fashion: the heuristic is needed to generate a set of

[2] Although Chen (2013) proposes an even faster (and reported better results on the BR corpus), his work was unpublished at the time we made our own experiments. Our work, published in (Magistry and Sagot, 2013) and presented in chapter 10 is an improvement over Zhikov et al.'s proposal. Chen does not provide results on any MSC dataset.

better segmentation candidate from a plausible segmentation.

Zhikov et al. (2010) propose two distinct iterative procedures that they combine sequentially. The first one operates on the whole corpus. They begin by ordering all possible word-boundary positions using BE and then try to add word boundaries checking each position sorted by decreasing BE, and to remove word boundaries checking each position by increasing order of BE. They accept any modification that will result in a smaller DL. The rationale behind this strategy is simple: for a given position, the higher the BE, the more likely it is to be a word-boundary. They process the more likely cases first. The main limitation of this procedure is that it is unable to change more than one position at a time. It will miss any optimisation that would require to change many occurrences of the same string, e.g., if the same mistake is repeated in many similar places, which is likely to happen given their initial segmentation algorithm.
To overcome this limitation, Zhikov et al. (2010) propose a second procedure that focuses on the lexicon rather than on the corpus. This procedure algorithm tries (i) to split each word of the lexicon (at each position within each word type) and reproduce this split on all occurrences of the word, and (ii) to merge all occurrences of each bi-gram in the corpus provided the merge results in an already existing word type. This strategy allows them to change multiple positions at the same time but their merging procedure is unable to discover new long types that are absent from the initial lexicon.

Our own implementation of Zhikov et al. (2010)'s system tested on MSC corpora yield the following F_w : 0.80, 0.79, 0.78 and 0.76 on PKU, CITYU, MSR and AS corpora, respectively. We will show in chapter 10 how we can outperform these results.

MDL can be related to Bayesian inference if we consider the lexicon M_w as the hyptohesis h. The objective is to minimize $L(C)$, thus to find:

$$\hat{M_w} = \arg\min_{M_w}(L(D|M_w) + L(M_w)),$$

$$\hat{M_w} = \arg\min_{M_w}(-log_2(P(D|M_w)) + L(M_w)),$$

$$\hat{M_w} = \arg\max_{M_w}(P(D|M_w) \times 2^{-L(M_w)}),$$

which is equivalent to a Bayesian model with a particular prior which states that the probability of a hypothesis is formulated as a function of the lexi-

con which decreases exponentially with the length of the lexicon (stating that $P(h) = f(M_w) \propto 2^{-L(M_w)}$).

7.2.5. Combining Systems

Some segmentation systems proposed in the literature combine multiple cues to make decision. For example, Zhao and Kit (2008) combine a compression based measure related to MDL to segment two-character words and AV to segment longer words.

Hewlett and Cohen (2009) propose to use Bootstrap Voting Expert to combine the *votes* of different indicators, two of which being based on BE (either high BE at boundaries for separation or low internal BE for cohesion). They use MDL to set various parameters in their model. In (Hewlett and Cohen, 2011) they combine this system with an MDL procedure to improve the segmentation.

Wang et al. (2011) present ESA, "Evaluation, Selection, Adjustment." This method combines cohesion and separation measures in a "goodness" metric that is maximized during an iterative process. The main drawbacks of ESA is the need to set a parameter that balances the impact of the cohesion measure w.r.t. the separation measure. Empirically, a correlation is found between the parameter and the size of the corpus but this correlation depends on the script used in the corpus (it changes if Latin letters and Arabic numbers are taken into account during pre-processing or not). Moreover, computing this correlation and finding the best value for the parameter (i.e., what the authors call the *proper exponent*) requires a manually segmented training corpus. Therefore, this proper exponent may not be easily available in all situations and the unsupervised nature of the approach is questionable.

7.2.6. Evaluating Unsupervised Segmentation Systems: Results and Issues

Evaluating unsupervised systems is a challenge by itself. As we discussed in the previous chapter, there is no clear definition of the linguistic objectives. On the other hand, corpora available for evaluation use heuristics to achieve consistency that are sometimes questionable and different corpora may disagree on various points. The evaluation of supervised systems can be achieved on any corpus using any guidelines: when trained on data that follows particular guidelines, the resulting system is supposed to follow these specific guidelines as well as possible, thus can it be evaluated on data annotated accordingly. However, with unsupervised learning

there is no reason why a system should be closer to one reference than another or even not to lie somewhere in between the different existing guidelines.

Huang and Zhao (2007) propose to use cross-training of a supervised segmentation system in order to have an estimation of the consistency between different segmentation guidelines. Zhao and Kit (2008) consider that these consistency levels provide a sensible upper bound of what can be expected from an unsupervized system. In their experiments, the average consistency is found to be as low as 0.848 (F_w).

Per word-length evaluation is also important as units of various lengths tend to have very different distributions. We also expect words of different length to roughly represent different cases of lexical formation that may have a strong influence on the consistency on the annotation. We used ZPAR (Zhang and Clark, 2010) on the four corpora of the Second Bakeoff to reproduce Huang and Zhao (2007) experiments and also measure the consistency per word length.

Our overall results are presented in Figures 7.2 to 7.5. Overall results are comparable to what Huang and Zhao (2007) report. However, consistency scores are quickly falling for longer words: on unigrams, f-scores range from 0.81 to 0.90 (the same as the overall results). We get slightly higher results on bigrams (0.85-0.92) but much lower on trigrams with only 0.59-0.79. We shall underline here that the lower results on longer words is likely to be the result of both lower consistency and sparser data.

Another issue about the evaluation and comparison of unsupervised systems is to try and remain fair in terms of pre-processing and prior knowledge given to the systems. In CWS, systems may especially be very sensitive to the way we deal with non-Chinese characters, some factoids can also be processed with specific rules. It can affect both the quality of the segmentation of those types and the quality of the estimation of relevant statistics about Chinese characters (to some extent, non-Chinese character can be considered as introducing "noise" in the models, as their distribution is likely to be very different). This should be kept in mind when comparing published figures from different sources. To be fairly comparable with other systems, we implemented our own versions of Jin's algorithm and Zhikov et al. algorithms. They will be used in the second part of this thesis.

In the literature, a large variety of stances are adopted. Hewlett and Cohen (2009) consider that the sentence segmentation in the BR corpus is a kind of supervision. But aiming at an fully unsupervised setup, they strip punctuation out of the Chinese corpus during the preparation of the data. This can be considered as preprocessing

based on knowledge about the script (although the result of this preprocessing is linguistically questionable). Wang et al. (2011) provide numerous results of their system with various preprocessing options (including or not specific treatment of punctuation, numbers and latin letters) and try to find the correct setting to compare their system with other published results. For reference, we still provide the results that were published for other systems in Table 7.6

		Entrainement			
		AS	CITY-U	PKU	MSR
test	AS	.945	.899	.843	.827
	CITY-U	.882	.951	.857	.810
	PKU	.866	.875	.948	.853
	MSR	.818	.826	.858	.969

Table 7.2.: F_w **obtained by a cross-trained ZPAR (Zhang & Clark 08), all words**

		Entrainement			
		AS	CITY-U	PKU	MSR
test	AS	.948	.901	.850	.836
	CITY-U	.877	.957	.858	.812
	PKU	.860	.866	.947	.872
	MSR	.831	.829	.872	.970

Table 7.3.: F_w **obtained by a cross-trained ZPAR (Zhang & Clark 08), unigrams only**

		Entrainement			
		AS	CITY-U	PKU	MSR
test	AS	.960	.923	.875	.865
	CITY-U	.910	.960	.884	.854
	PKU	.902	.905	.961	.886
	MSR	.854	.867	.890	.977

Table 7.4.: F_w **obtained by a cross-trained ZPAR (Zhang & Clark 08), bigrams only**

		Entrainement			
		AS	CITY-U	PKU	MSR
test	AS	.866	.779	.620	.573
	CITY-U	.790	.900	.723	.644
	PKU	.695	.758	.887	.595
	MSR	.604	.632	.593	.935

Table 7.5.: F_w **obtained by a cross-trained ZPAR (Zhang & Clark 08), trigrams only**

	Bakeoff2 (with preproc)				Bakeoff3 (without preproc)				BR corpus
System	PKU	MSR	CITYU	AS	CTB	MSR	CITYU	AS	BR
DLG+AV					.658	.667	.692	.663	
Jin	.643	.598	.573	.609					
Zhikov	.808	.782	.787	.762					
Wang et al.		.819	.828		.770	.760	.757	.752	
Mochiachi et al.		.807	.824						.757

Table 7.6.: **Reported results of various segmentation systems**

Part II.

Autonomy Measure and Segmentation Algorithms

Chapter 8

Variations on the Harrissian Hypothesis

8.1. The Choice of the Harrissian Hypothesis

In the first part of this thesis, we underlined the need to define an objective way of measuring the *autonomy* of a form as the first step towards a computable definition of wordhood.

We introduced a collection of unsupervised segmentation systems from which we draw our inspiration and the methodology to evaluate the quality of the resulting segmentation. Although it is not perfect, this evaluation methodology will provide a good starting point to ensure the relevance of our own proposal.

The Harrissian hypothesis and its reformulation by Kempe (1999) using the Variation of Branching Entropy as well as the reported results on Chinese Word Segmentation by Jin and Tanaka-Ishii (2006) are especially appealing for our own goal. However, they focus on taking very local decisions at the word boundary level. In this chapter, we show how it is possible to define a measure of the *autonomy* of any sequences of characters based on the same clues. This measure can in turn be used to design a segmentation algorithm for CWS that improves on all other systems based on BE or VBE. With the enhancements of the following chapters, we reach state-of-the-art scores on unsupervised CWS.

8.2. From Elicitation to Corpus-Based Estimation

At the core of the Harrissian hypothesis is the level of uncertainty about what may follow or precede a given sequence of symbols. When reading from left to right, as we read more and more characters and thus obtain more and more information, one can expect a decreasing uncertainty about the next character. But when one reaches the end of a word, the level of uncertainty is expected to suddenly rise again.

If we are given a way to estimate the probability of every possible following character given a string, *entropy* provides a sensible measure of the uncertainty.

8.2.1. Formulation

Let us now recall and detail the definitions of BE and VBE, already sketched in chapter 7, which can be computed by reading the corpus forward (to obtain the values for the right contexts of each form) or backward (to obtain the values for the left contexts of each form).

Given an n-gram $x_{0..n} = x_{0..1}x_{1..2}\dots x_{n-1..n}$ (where the indices are indexing the positions between every two characters) with a right context $\chi_{\rightarrow}$, we define its *Right Branching Entropy* (RBE) as:

$$\begin{aligned} h_{\rightarrow}(x_{0..n}) &= H(\chi_{\rightarrow} \mid x_{0..n}) \\ &= -\sum_{x \in \chi_{\rightarrow}} P(x \mid x_{0..n}) \log P(x \mid x_{0..n}). \end{aligned}$$

The *Left Branching Entropy* (LBE) is defined in a symmetric way: if we note $\chi_{\leftarrow}$ the left context of $x_{0..n}$, its LBE is defined as:

$$h_{\leftarrow}(x_{0..n}) = H(\chi_{\leftarrow} \mid x_{0..n}).$$

The RBE $h_{\rightarrow}(x_{0..n})$ can be considered as $x_{0..n}$'s *Branching Entropy* (BE) when reading from left to right, whereas the LBE is $x_{0..n}$'s BE when reading from right to left.

From $h_{\rightarrow}(x_{0..n})$ and $h_{\rightarrow}(x_{0..n-1})$ on the one hand, and from $h_{\leftarrow}(x_{0..n})$ and $h_{\leftarrow}(x_{1..n})$ on the other hand, we estimate the *V*ariation of Branching Entropy (VBE) in both

directions, defined as follows:

$$\begin{aligned}\delta h_{\rightarrow}(x_{0..n}) &= h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1}) \\ \delta h_{\leftarrow}(x_{0..n}) &= h_{\leftarrow}(x_{0..n}) - h_{\leftarrow}(x_{1..n}).\end{aligned}$$

With these definitions, all we need is a way to estimate $P(c|x_{0..n})$ to obtain our measure of uncertainty and its variation. This is precisely the role of *Language Models*. In this work, we simply estimate this probability based on the count of occurrences in the corpus (just as Kempe (1999); Tanaka-Ishii (2005); Jin and Tanaka-Ishii (2006) do). $P(c|x_{0..n})$ is given by the ratio of the number of occurrences of the context followed (resp. preceded) by c on the number of occurrences of the context with or without c.

$$P(c|x_{0..n}) = \frac{\#x_{0..n+1}}{\#x_{0..n}}, \text{ with } x_{n..n+1} = c$$

8.2.2. How Probability Estimation Differs from Elicitation

Although the branching entropy of a Language Model seems to be a sensible estimation of contextual uncertainty for a speaker, it presents some important differences with the initial methodology proposed by Harris.

Figures obtained from elicitation rely on the faculty the speaker has to generate a potentially infinite number of utterances. Probabilities on the other hand are estimated from a finite corpus. The inherently finite size of the corpus results in a *data sparsity* problem that grows with the length of the context considered. Within a large corpus, we can expect to observe all the usual characters. We may observe many short strings, but we will see only a little proportion of all the sentences that a native speaker could think of.

When asked how many phonemes can follow *[hiyzkwi]* (*he is qui-*) , a native speaker can generate new utterances "on demand". This is impossible for a study on a corpus. The full utterance *[hiyzkwikər]* may be absent from the corpus, even if *[hi]*, *[yz]* and *[kwikər]* are all present many times in other contexts in the corpus.

To overcome this issue, the usual solution is to use either larger corpora or smaller context. Kempe (1999) uses a 3-characters sliding window. So that the context is always of length three. This means that he makes the following approximation $H(\chi_{\rightarrow} \mid [\text{hiyzkwi}]) \approx H(\chi_{\rightarrow} \mid [\text{kwi}])$. Jin and Tanaka-Ishii (2006) on the other hand use multiple language models of order 1 to 6 and retain the higher variation.

Another noteworthy distinction between Kempe's and Jin's systems is that Jin uses a variable context size where Kempe uses a constant one. Therefore Jin's VBE is similar to our definition but Kempe's VBE differs slightly. For the same example, in the utterance *[hiyzkwikər]*, reading from left to right, Kempe will compute $VBE_{\rightarrow}([\text{kwi}]) = H(\chi_{\rightarrow} \mid [\text{kwi}]) - H(\chi_{\rightarrow} \mid [\text{zkw}])$. Jin will compute $VBE_{\rightarrow}([\text{kwi}]) = H(\chi_{\rightarrow} \mid [\text{kwi}]) - H(\chi_{\rightarrow} \mid [\text{kw}])$ (So the computation of the VBE requires language models based on two different contexts length).

We prefer Jin's formulation as it can be expected to behave more like the evolution of uncertainty described by Harris. In the general case, due to data sparsity, the BE of a longer string is expected to be lower than the BE of the shorter string. Therefore, the VBE is decreasing in general, an increasing VBE (or even a relatively less decreasing VBE) seems more meaningful under Tanaka-ishii and Jin's formulation than it is under Kempe's.

One of the main distinctions between our system and other systems based on VBE or BE is that we want to focus on measuring a property of the word candidates. All other systems aim at affecting a binary value to the candidate boundaries. This allows us to deal with data sparsity more elegantly and robustly.

For example, Jin's system will ask whether the position after *[hiyzkwik]* is a boundary or not. To do so, he will compute the VBE of *[yzkwik]*, *[zkwik]*, *[kwik]*, *[wik]*, *[ik]* and *[k]* and consider the higher variation (strictly speaking he also does so for the 6 contexts at the right of the considered position, but reading from right to left and it will take the maximum of the 12 values).

Our system rather asks whether *[kwik]* is an autonomous wordform or not. To do so, we only have to consider $VBE_{\rightarrow}([\text{kwik}])$ and $VBE_{\leftarrow}([\text{kwik}])$. A high autonomy value will mean that a boundary at the right of *[kwik]* is expected when we do the estimation of VBE starting from its left and a boundary at the left of it is expected when we do the estimation starting from its right. We will see that this simpler starting point allows us to reach much higher segmentation scores by taking our decisions more globally, at the sentence level.

Although our method of computation frees us from the need to start at the beginning of each utterance, we still have to consider closely what can be the effects of *data sparsity* on our computation, especially for the likely case of long sequences.

We work in an unsupervised fashion that allows us to do *endogenous* learning. This means that the data we have to segment is a part of our training data. Therefore

we will never have to deal with unseen data (a major issue in supervised learning). Nevertheless, as we are expecting the occurrences of the forms in our corpora to follow a Zipfian (Zipf, 1949) distribution, we will encounter a large number of hapax legomena (*i.e.*, forms that occur only once in the corpus).

If we consider sequences of characters of increasing lengths (by taking the previous sequence and appending a character), the number of occurrences observed in the corpus will drop until we reach a unique occurrence.

We can easily show that any position in the corpus will always be included in a hapax legomenon if we consider a sequence large enough: all the positions are included in the corpus and the corpus itself is a sequence that occurs once.

Let us now consider the RBE of a hapax. As the hapax occurs only one time in the corpus, it will have only one possible following character. This means that we reach certainty and its RBE will be 0. The same can be said of its LBE.

Now if we consider the BEs of a larger hapax by one character, it will obviously be a hapax too. Therefore the VBE inside a hapax will be 0, *i.e.*, non-decreasing and suspiciously high. This case of constant BE at 0 should not receive the same interpretation as the usual case of a constant non 0 BE (in which the VBE will be 0). It is a sign of *data sparsity*. We shall not consider such sequences as word candidates but as a lack of data.

As a consequence, once the BE as fallen to 0, we do not consider the larger strings with a null BE as word candidates with valid VBE value. We simply leave them out.

Note that it may be the case that a n-gram is not an hapax but always appears in the same context. Its internal BEs will drop to 0 and eventually (if it is long enough), its internal VBE will do too. All the values will stay at 0 until we reach the position where the larger n-gram occurs in different contexts, then the BE will rise again and we shall conclude that a word boundary is very likely.

This case will actually correspond to multi-characters bound morphemes and rare words. Discarding the hypothesis that such sequences could be splitted into multiple words is a sensible way to model our hypothesis about *autonomy*. It is thus reasonable to apply the same rule when the BE is and stays null even if it is not a case of data sparsity. In other words, multi-characters bound morphemes can be frequent items in a corpus, but we don't have to try to split them if we have no evidence of alternation. They will not be associated to a low autonomy value, they rather will have no autonomy at all.

We thus have to slightly change the definitions of VBEs:

$$\delta h_{\rightarrow}(x_{0..n}) = \begin{cases} h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1}), & \text{if } h_{\rightarrow}(x_{0..n}) \neq 0 \text{ or } h_{\rightarrow}(x_{0..n-1}) \neq 0 \\ \text{undefined} & \text{otherwise} \end{cases}$$

$$\delta h_{\leftarrow}(x_{0..n}) = \begin{cases} h_{\leftarrow}(x_{0..n}) - h_{\leftarrow}(x_{1..n}), & \text{if } h_{\leftarrow}(x_{0..n}) \neq 0 \text{ or } h_{\leftarrow}(x_{1..n}) \neq 0 \\ \text{undefined} & \text{otherwise} \end{cases}$$

8.2.3. VBE as a Cue for Wordhood

Let us now illustrate the relevance of our approach on the available data. We shall now confirm that Harris hypothesis and our way to model it provide valuable insights for our goal.

We will first limit the discussion to the $VBE_{\rightarrow}$. Figures 8.1 to 8.6 show the probability density functions of the values that the $VBE_{\rightarrow}$ can take for different lengths of n-grams. It can be read as an estimation of the probability to find a certain value of VBE given the length of an observed n-gram. Each figure shows three curves: one computed on all the n-grams, one computed by taking only the n-grams that are segmented as a word at least once in the manually segmented corpus and one computed by taking only the n-grams that are never segmented as a word in the corpus.

These figures show that manually segmented words in our corpus are more likely to have a higher VBE than non-words and that they form a small proportion of the observed strings (this proportion decreases as the length of the strings increase). However, there still is a large area under both curves where a boundary decision seems hard to make based solely on a single VBE value.

We can make further observations with the same method. One important question is whether the VBE inside the word candidates is significantly different from non-word and shall be used in our inference as well or if we can rely solely on the positions at the beginning and at the end of each form. To answer this question, we run the same kind of density distribution estimation but on the various possible values at the boundaries and at the inner position of words and non words of a fixed length. We illustrate this with trigrams in Figure 8.7. Considering all the trigrams ABC, we compute the VBEs for all prefixes (reading from left to right) and all suffixes (reading from right to left) and plot the probability density function based

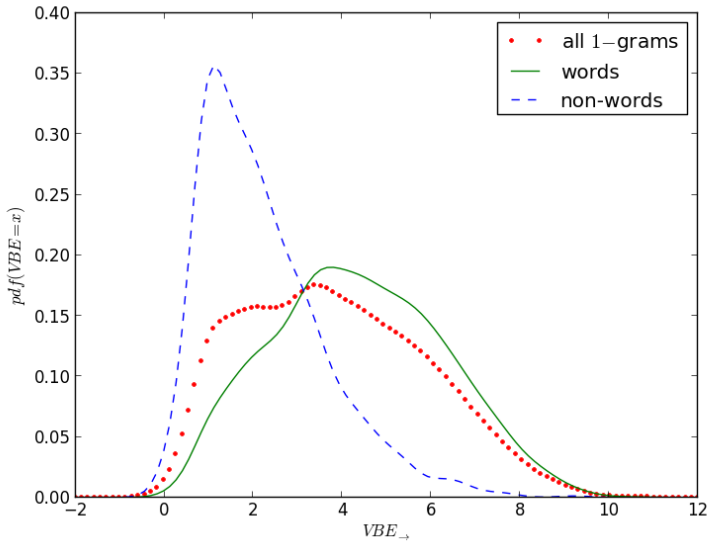


Figure 8.1.: **Probability Density functions of $VBE_{\rightarrow}$ values for 1-grams**

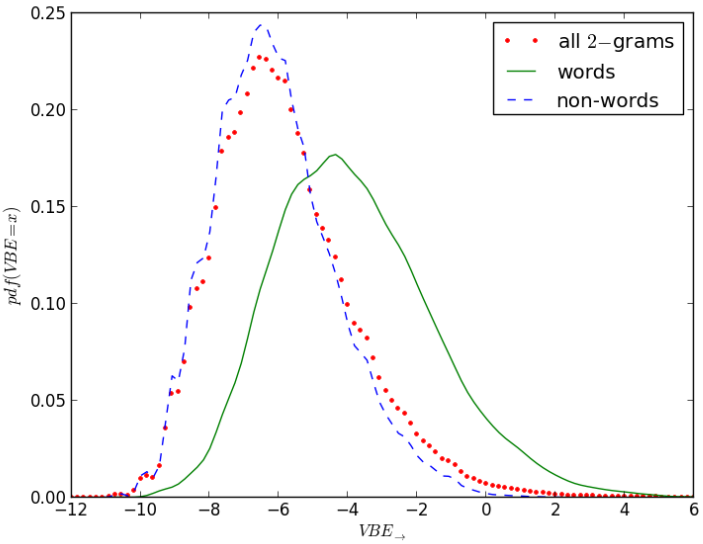


Figure 8.2.: **Probability density functions of $VBE_{\rightarrow}$ values for 2-grams**

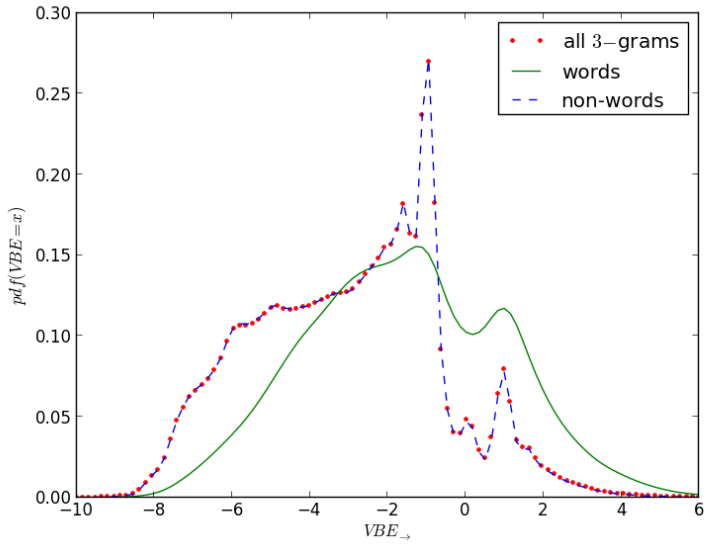


Figure 8.3.: **Probability density functions of $VBE_{\rightarrow}$ values for 3-grams**

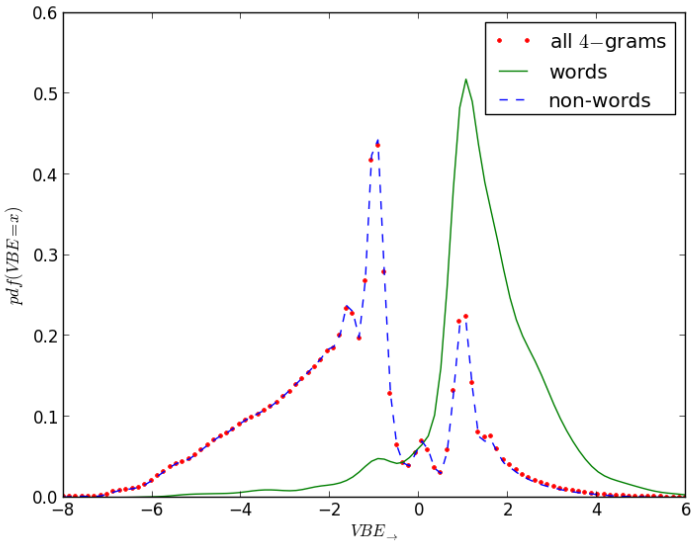


Figure 8.4.: **Probability density functions of $VBE_{\rightarrow}$ values for 4-grams**

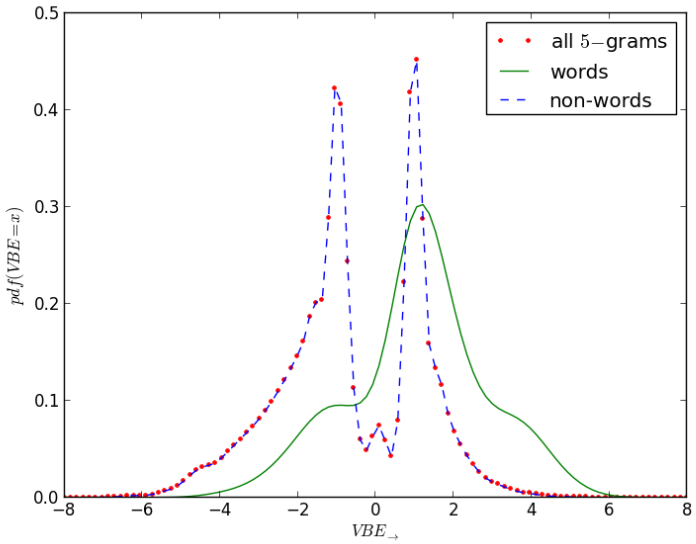


Figure 8.5.: **Probability density functions of $VBE_{\rightarrow}$ values for 5-grams**

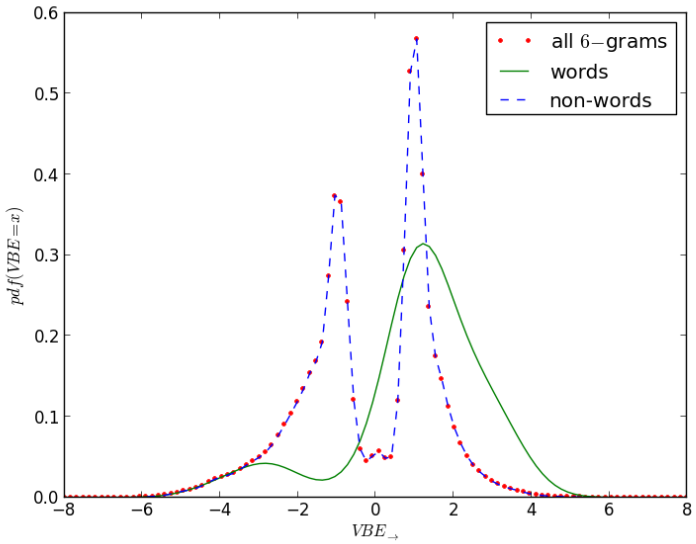


Figure 8.6.: **Probability Density functions of $VBE_{\rightarrow}$ values for 6-grams**

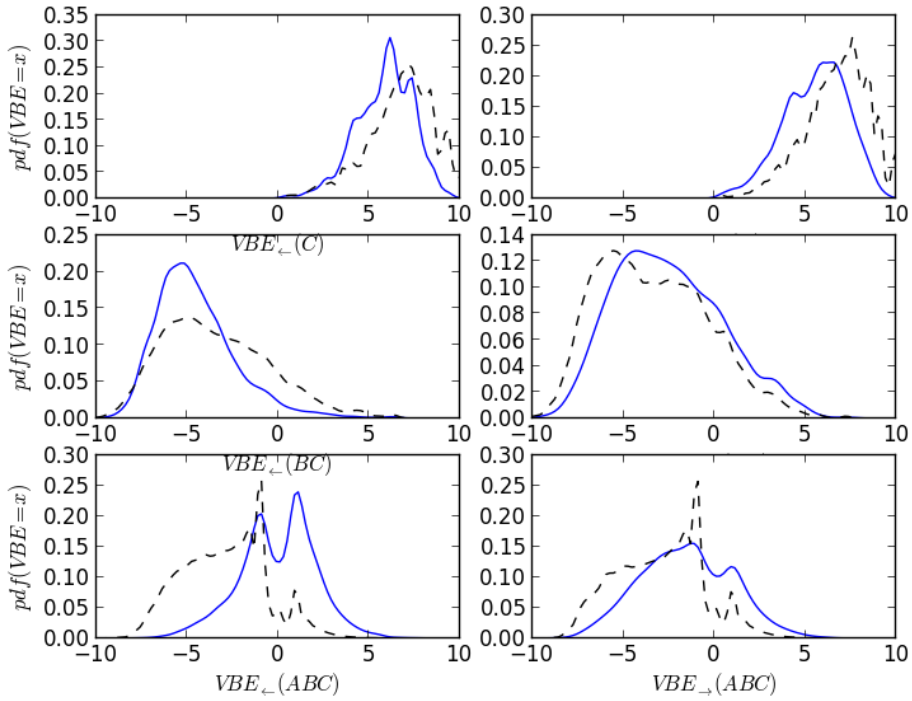


Figure 8.7.: **VBE of all substrings of the trigrams (dashed black) and for the substrings of the trigrams that are segmented as words in the manually segmented corpus (blue line).**

on the data of all the *n*-grams of the given length (in dashed black) and based only on the *n*-grams that are prefixes or suffixes of an attested word of length 3 in the manual segmentation.

What we observe is that the values with the higher discriminative power seem to be $VBE \rightarrow (ABC)$ and $VBE \leftarrow (ABC)$, that is to say the two word boundaries after having read a whole word. Although we expect the *BE* to decrease inside a wordform as we read it, we are apparently unable to distinguish words from non-words based on a more specific decrease at inner positions. It may be so that such values seem more likely to be affected by morphology or that this decrease is not distinguishable from the decrease of random data. In any case, we shall focus on the distinguishable increase of VBE at border position of wordforms to achieve our goal.

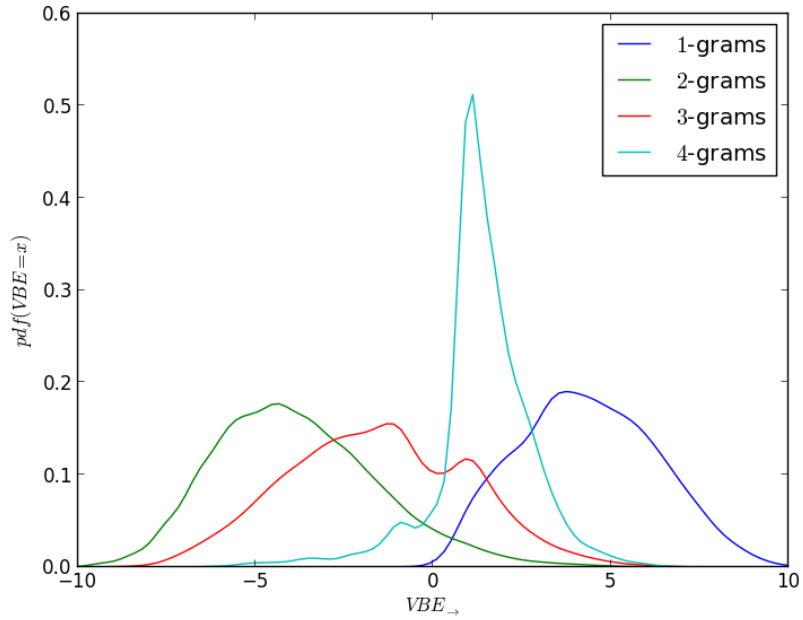


Figure 8.8.: $VBE_{\rightarrow}$ **for words of different lengths.**

8.2.4. The Need for Normalisation

In the general case or on random data, we expect the BE to decrease as we read a longer string. We showed that consistently with Harris Hypothesis, this general tendency does not hold when reaching a word boundary. The observations we just described showed that the values of the VBE at the begining and at the end of strings is able to help us to distinguish words from non-words among n-grams of a given length.

However, nothing allows us to conclude that this distinction can be made if we measure the VBE for n-grams of different lengths.

We thus tried to observe the values taken by the VBE at the boundaries of words for different values of n. These observations are reported for $1 \geq n \geq 4$ on Figure 8.8.

This figure underlines the need for a normalisation procedure before VBE values can be compared for n-grams of different lengths.

8.2.5. Consequences for the Autonomy Measure

We just observed that

- the *VBE* computed at the extremities of word candidates, from the opposite extremity of each *n*-gram are the most relevant for our purpose;
- as there is some overlap between words and non-words, local values will not be enough to perform segmentation and some global optimisation is needed;
- to compare these values for *n*-grams of different lengths, some normalisation is needed.

These points have important consequences for our segmentation system. The relevant values to judge the *autonomy* of a *n*-gram $x_{0..n}$ are $VBE_{\rightarrow}(x_{0..n})$ and $VBE_{\leftarrow}(x_{0..n})$. The *autonomy* we want to define will thus be a function of these two values. These two values have the interesting property that although they are computed based on all the occurences of $x_{0..n}$, $x_{0..n-1}$ and $x_{1..n}$, once computed they are independent of any specific context. This enable us to precompute the involved values of VBE for each possible *n*-gram and to perform the required normalisation.

8.3. A Novel Segmentation Algorithm

As we saw in the previous section, normalisation, combination of $VBE_{\rightarrow}$ and $VBE_{\leftarrow}$ and global optimisation are required to propose a sensible segmentation algorithm based on *VBE* as an *autonomy* estimate.

8.3.1. Normalisation

For normalisation, we tried to center the values around 0 by simply substracting the average *VBE* of all strings of a given length from the *VBE*s of each string of this length. Another possibility is to compute a z-score (*i.e.*, to *standardise* the value) to also control the spread of the values. We note *NVBE* the Normalised Variation of Branching Entropy.

For each length k and each k-gram x such as $len(x) = k$, with μ_k the mean of all the values $VBE(y)$ such as $len(y) = k$, we can define a first variant of *NVBE* as:

$$N_{\mu}VBE(x) = VBE(x) - \mu_k.$$

We define in a symmetric way $N_\mu VBE_\leftarrow$ and $N_\mu VBE_\rightarrow$.

Using the z-score function, we define an other variant of $NVBE$ as follow.

$$N_z VBE(x) = \frac{VBE(x) - \mu_k}{\sigma_k},$$

Where σ_k is the standard deviation of all the values $VBE(y)$ such as $len(y) = k$.

In the remaining of the dissertation, we use $NVBE$ to denote either $N_\mu VBE$ or $N_z VBE$.

8.3.2. Autonomy Function

Once we have obtained normalised values $NVBE$s, we can introduce an *autonomy* function $a(x_{0..n})$ which associates a single autonomy value to any n-gram $x_{0..n}$ based on $NVBE_\rightarrow(x_{0..n})$ and $NVBE_\leftarrow(x_{0..n})$. We considered two simple functions:

- the sum of the two values
- the minimum of the two values

In both cases, the higher $a(x)$, the more likely is x autonomous.

$a(x)$ is a single value independent of a specific context in which x may occur. It can be pre-computed for each n-gram x in the corpus.

8.3.3. Segmentation Algorithm

With this measure, we can redefine the problem of the segmentation of a sentence as the maximization of the autonomy measure of its words. For a character sequence s, if we call $Seg(s)$ the set of all the possible segmentations, then we are looking for :

$$\underset{W \in \mathrm{Seg}(s)}{\arg\max} \sum_{w_i \in W} a(w_i) \cdot \mathrm{len}(w_i)$$

where W is the segmentation corresponding to the word sequence $w_0 w_1 \ldots w_m$ and $len(w_i)$ is the length of a word w_i used here to be able to compare segmentations resulting in a different number of words. This best segmentation can be computed easily using dynamic programming.

8.3.4. Dynamic Programming Formulation

For a given string $u_{0..k}$ of length k, there are 2^{k-1} possible segmentations. However, we shall note that if we already know the best segmentation for $u_{0..k}$ and for each of its prefixes $u_{0..n}$, $n \leq k$, segmenting the string $u_{0..k+1}$ made of $u_{0..k}$ and the following character only requires to consider to which "word" the following character should belong. Because we are limiting ourselves to continuous units, we only have to consider the following cases:

- The $k+1$th character is a word of length 1 that we just append at the end of the best segmentation of $u_{0..k}$.
- for each prefix $u_{0..n}$ of $u_{0..k}$ $(0 < n < k)$, $u_{n..k+1}$ is a word that we append to the best segmentation of $u_{0..n}$.
- the whole string $u_{0..k+1}$ is a single word

The first and last cases can be considered as extreme cases for the middle case if we take $0 \leq n \leq k$. They are explicitly mentioned here for the sake of clarity.

This allows us to define the best segmentation of $u_{0..k+1}$ recursively from the best segmentations of its prefixes:

$$\underset{W \in \mathrm{Seg}(u_{0..k+1})}{\arg\max} = \underset{V \in \bigcup_{n \leq k} \left(\bigcup_{S \in \mathrm{Seg}(u_{0..n})} S \cup \{u_{n..k+1}\} \right)}{\arg\max} \sum_{w_i \in V} a(w_i) \cdot \mathrm{len}(w_i)$$

This recursive formulation allows us to use dynamic programming. We just have to keep in memory the best segmentation at each step $\mathrm{Seg}(u_{0..n})$ so we consider only $\sum_{n=2}^{k} n$ segmentations rather than 2^{k-1}.

This Viterbi decoding algorithm thus have a quadratic run time rather than an exponential one. It results in a light overhead when $k < 5$ but is more efficient if $k \geq 5$ and increasingly interesting as k grows.

8.4. Quantitative Results

We have implemented the segmentation algorithm with the two variants of *NVBE* as well as Jin's and Zhikov *et al.*'s algorithms for the sake of comparison. We ran an evaluation on the four corpora which were made available for the Chinese Word Segmentation Bakeoff 2 described in chapter 6. Results are shown in Table 8.1.

These results show only a little difference between our two normalisation procedures. We are still below the *state-of-the-art* of the systems that combine multiple types of clues but we are already achieving a better segmentation than other systems strictly based on *VBE*.

Considering the simplicity of our algorithm and our goal to obtain easily interpretable results for further linguistic analysis, we see these results as very promising.

Beside the overall higher scores of our system, it is worth noting that it is less sensitive to the maximum order of the language models than the other systems. Increasing the maximum length of the considered n-grams only affects the computation time in our case but dramatically affects our implementation of Jin's system. On the other hand, Zhikov's initialisation includes an unsupervised way to automatically select the order of the language model. We shall also stress that Zhikov *et al.* do not intend to use their initialisation procedure "as-is" but only as the first step for their MDL procedure (see chapter 10).

Another noteworthy difference concerns the boundary-base recall (R_b) and precision (P_b) scores. For our system, $R_b > P_b$ where for others $P_b > R_b$. This means that we tend to over-segment the input where others tends to undersegment. We believe it is a good thing as we tend to identify units that may be autonomous in other contexts, to conclude that they are not free in a given context requires insight of a different kind (related to word classes) that is not addressed yet. This is consistent with our linguistic discussion. It also has important consequences when augmenting such systems with a MDL optimisation procedure.

Method	F_w	F_b	R_b	P_b
PKU corpus				
Jin ($n \leq 3$)	0.643	0.849	0.783	0.926
Zhikov *et* al. initialisation ($n = 2$)	0.584	0.802	0.681	**0.974**
$N_\mu VBE$, min	0.750	0.905	0.917	0.893
$N_z VBE$, min	0.737	0.899	0.903	0.895
$N_\mu VBE$, sum	**0.761**	**0.910**	**0.923**	0.896
$N_z VBE$, sum	0.758	0.908	0.918	0.890
City-U corpus				
Jin	0.573	0.814	0.732	0.917
Zhikov *et* al. initialisation	0.534	0.779	0.647	**0.981**
$N_\mu VBE$, min	0.715	0.895	0.929	0.863
$N_z VBE$, min	0.712	0.894	0.918	0.870
$N_\mu VBE$, sum	0.731	**0.901**	**0.937**	0.867
$N_z VBE$, sum	**0.733**	**0.901**	0.928	0.876
MSR corpus				
Jin	0.598	0.841	0.907	0.783
Zhikov *et* al. initialisation	0.624	0.835	0.740	**0.957**
$N_\mu VBE$, min	0.766	0.911	0.936	0.888
$N_z VBE$, min	0.762	0.909	0.925	0.893
$N_\mu VBE$, sum	0.786	**0.919**	**0.946**	0.894
$N_z VBE$, sum	**0.787**	**0.919**	0.940	0.899
AS corpus				
Jin	0.609	0.847	0.860	0.834
Zhikov *et* al. initialisation	0.540	0.809	0.688	**0.981**
$N_\mu VBE$, min	0.718	0.906	0.928	0.884
$N_z VBE$, min	0.719	0.905	0.919	0.892
$N_\mu VBE$, sum	0.734	0.911	**0.932**	0.891
$N_z VBE$, sum	**0.745**	**0.914**	0.928	0.900

Table 8.1.: **Results of the different segmentation algorithms inspired by the Harrissian hypothesis on the Bakeoff 2 testset.**

Chapter 9

Rationale and Effects of preprocessing

We saw in chapter 6 that an important part of discrepancies between the various segmentation guidelines concerns the so-called "factoids". This term covers a variety of language phenomena that include: numbers, dates, addresses, email addresses, proper names… As we shall now argue, a specific treatment of a subset of such expressions is both sound and efficient.

9.1. Refining the Definition and Processing of Factoids

Factoids are diverse but most of them share common properties that call for a specific treatment.

They are built along precise and often explicitly defined rules. These rules make their internal structure quite different from other language constructions. In fact, most of these rules are learnt through explicit teaching at school. In that respect, they arguably lie outside the scope of "natural" language and receive little attention from linguistic studies. Nevertheless, they are pervasive in language data that NLP systems have to process and in the data we want to draw our inferences from.

Their internal structure is characterized by a high degree of regularity and can often be described with formal grammars with a low complexity. This has two

important consequences: i) they can easily be dealt with using small sets of rules or regular expressions and ii) they are likely to introduce noise into our models as there is no reason why they should follow Harris's hypothesis.

For these reasons, a separate treatment of such expressions by adding a pre-processing step is both sound regarding the initial hypothesis that we try to model and promising for the performance of our system.

We thus consider that we can and should pre-process any expression that

- is arguably outside the scope of "natural" language for its construction rules are typically explicitely learnt (for example, how to read a watch to tell or write what time it is).
- can be captured by a simple set of regular expressions (that we call a "local grammar")

In order to do so in a generic and reusable fashion, we decided to extend the preprocessing pipepline Sxpipe (Sagot and Boullier, 2005) initially developed for French but which was already extended to a variety of languages. We augment its local grammars for basic pre-processing of MSC factoids.

It is worth recalling that these expressions are an important area of disagreement between the various segmentation guidelines (chapter 6). Their "gold" segmentation is even more arbitrary than for the more spontaneously created expressions.

In the following sections, we describe the local grammars we use and evaluate our system after their application.

9.2. Numbers

Numbers are quite straightforward to identify. They can be made of digits from the sets of Arabic numerals or Chinese numerals. A particularity of MSC related to numbers is the availability of a set of characters standing for various powers of 10 that can be combined with the digits. For example, 12,000 can also be written 1萬2 or even 1.2 萬 , with 萬 (wàn) standing for 10,000.

As the digits can eventually be used to create non-numeric words, we do not mark as *NUMBER* the isolated digit and treat them as normal characters.

Numerals can be turned into ordinals by adding a prefix 第 dì.

The regular expressions used to process numbers are reported in Figure 9.1.

Name	Regex
DIGIT	[○零一兩二三四五六七八九十〇 0-9]
P10	[十百千萬億兆]
PNT	[. ． 點]
NUMBER	-? DIGIT (DIGIT?P10?)+ (PNT (DIGIT?P10)+)?
ORDINAL	第 DIGIT | 第 NUMBER
PROPORTION	(DIGIT| 百) 分之 NUMBER

Table 9.1.: **Regular expressions for numbers used in preprocessing.**

9.3. Date and time

Date and time expressions are a succession of numeric expressions and temporal units such as 年 nián 'year', 月 yuè 'month', 日 rì 'day' (or 號 hào 'number'), 時 shí 'hour', 分 fēn 'minute' (or 分鐘 fēnzhōng) and 秒 miǎo 'second').

A closed set of expressions can also be used to specify the period of the day: 凌晨 língchén 'dawn', 早上 zǎoshang 'early morning', 上午 shàngwǔ 'morning', 中午 zhōngwǔ 'noon', 下午 xiàwǔ 'afternoon', 晚上 wǎnshàng 'evening'.

The regular expressions used for preprocessing the date and time expressions are presented in Table 9.2

9.4. Addresses

Addresses in Chinese are written with various location elements ordered in decreasing size: province, county, department, city/village, district, street, section, alley, lane, number, floor. (respectively: 省,縣,鄉,部,市,村,區,街/路/大道,段,巷,弄,號,樓). These elements are preceded by either a number (3rd floor) or a name.

Name	Regex
NUM12	DIGIT | 十[一二]? | 1[012])
NUM24	十? DIGIT | 1?[0-9] | 二十[一二三四]? | 2[0-4]
NUM31	[12]? [0-9] | 二?十? DIGIT | 三十一? ? | 3[01]
NUM60	[1-5]? [0-9] |[二三四五]? ? DIGIT
YEAR	(DIGIT| 百){1,4} 年)
MONTH	NUM12 月
DAY	NUM31 [日號]
HOUR	NUM24 (點|點鐘|时)
MINUTE	NUM60 (分钟|分鐘|分)
SECOND	NUM60 秒
PERIOD	凌晨|早上|上午|中午|下午|晚上
DATE	YEAR? MONTH? DAY?
TIME	PERIOD? HOUR? MINUTE? SECOND?

Table 9.2.: **Date and Time regular expressions used in preprocessing**

As the name for cities, streets, etc. are an open class, we allow any sequence of one or two characters but only consider as factoids the expression composed by a sequence of two elements or more (for example a city name followed by a district name 臺北市 萬華區 ‘Taipei city, Wanhua District’).
táiběishì wànhuáqū

This leads to the following pattern: (..?省)? (..?縣)? (..?鄉)? (..?部)? (..?市|..?村)? (..?區)? (..?街|路|大道) (NUMBER 段)? (NUMBER 巷)? (NUMBER 弄)? (NUMBER 號)? (NUMBER 樓)?

9.5. Web-related Factoids

URLs and emails are dealt with rules that were already included in Sxpipe.

9.6. Adding MSC to Sxpipe

We augmented the local grammars of Sxpipe to deal with the aforementioned expressions. This allow us to discard the matched expressions from the training data and segment them accordingly to the guidelines as a post-processing step.

Note that the regular expressions presented in the tables of this chapter were simplified for the sake of clarity. The actual implementation has to deal with the various ways to encode the same characters with different unicode codepoints, especially for numbers and Latin letters that are presents in both the Latin plane and the Chinese-related planes (where they are monospaced). Some of the characters presented here in traditional Chinese also have simplified Chinese counterparts in our grammars.

9.7. Quantitative Results

These results show a significant improvement overall. This confirms the soundness of a separate treatment of these phenomena. A rule-based system performs better on such expressions that strictly follow simple patterns and at the same time, it discards irrelevant data from the corpus for the estimation of the statistics to process the more natural part of the language.

Method	F_w	F_b	R_b	P_b
PKU corpus				
Jin	0.643	0.849	0.783	0.926
Jin, with sxpipe	0.675	0.868	0.826	0.916
Zhikov *et* al. initialisation	0.584	0.802	0.681	**0.974**
$N_\mu VBE$	0.761	0.910	0.923	0.896
$N_\mu VBE$, with sxpipe	**0.803**	**0.929**	**0.942**	0.916
$N_z VBE$	0.758	0.908	0.918	0.89
$N_z VBE$, with sxpipe	0.798	0.927	0.933	0.920
City-U corpus				
Jin	0.573	0.814	0.732	0.917
Jin, with sxpipe	0.603	0.837	0.790	0.890
Zhikov *et* al. initialisation	0.534	0.779	0.647	**0.981**
$N_\mu VBE$	0.731	0.901	0.937	0.867
$N_\mu VBE$, with sxpipe	0.772	0.918	**0.945**	0.892
$N_z VBE$	0.733	0.901	0.928	0.876
$N_z VBE$, with sxpipe	**0.774**	**0.919**	0.939	0.900
MSR corpus				
Jin	0.598	0.841	0.907	0.783
Jin, with sxpipe	0.610	0.848	0.906	0.798
Zhikov *et* al. initialisation	0.624	0.835	0.740	**0.957**
$N_\mu VBE$	0.786	0.919	0.946	0.894
$N_\mu VBE$, with sxpipe	0.790	**0.923**	**0.941**	0.905
$N_z VBE$	0.787	0.919	0.940	0.899
$N_z VBE$, with sxpipe	**0.793**	**0.923**	0.937	0.910
AS corpus				
Jin	0.609	0.847	0.860	0.834
Zhikov *et* al. initialisation	0.540	0.809	0.688	**0.981**
$N_\mu VBE$	0.734	0.911	0.932	0.891
$N_\mu VBE$, with sxpipe	0.759	0.921	0.921	0.921
$N_z VBE$	0.745	0.914	0.928	0.900
$N_z VBE$, with sxpipe	**0.764**	**0.922**	0.919	0.926

Table 9.3.: **Effects of pre-processing**

Chapter 10

Enhancement Using Minimum Description Length

We already presented the use of Minimum description Length (MDL) in various works on word segmentation in chapter 7 and related it to Bayesian inference. Especially, we described the work by Zhikov et al. (2010). We shall now see how their proposal can be combined with ours. We adapt their methodology to our definition of *autonomy* and our initial segmentation algorithm presented in the previous two chapters. Our definitions allow for a larger hypothesis search space that leads to important findings regarding the use of MDL in word segmentation.

Contrary to the widespread idea (Yu et al., 2002; Hewlett and Cohen, 2011) that a higher compression rate, i.e., a shorter Description Length, should correspond to a better segmentation, we shall show that as we are exploring a larger search space, we can reach a solution in which a lower DL corresponds to a lower segmentation F-score.

If we limit our search to linguistically more plausible hypotheses with a small set of rules resembling to some of the heuristics from the segmentation guidelines, we can reach state-of-the-art results.

The idea that the MDL of a unigram model can account for the quality of the segmentation is challenged by works on word segmentation using Bayesian inference. As we mentioned in section 7.2.2, the objective of the MDL is similar to the unigram model (DP) in (Goldwater et al., 2009) which yield unsatisfying results. We can argue that our MDL with a larger search space enables us to find solutions that are closer to those found by the DP model. We thus run into the same issue: we overfit a unigram model which is inappropriate for our linguistic purpose. Some constraints

must be added to avoid overfitting.

10.1. A New Segmentation Algorithm Based on MDL and NVBE

To enhance our segmentation system with a MDL procedure, we follow the general idea presented in (Zhikov et al., 2010).

Recall from chapter 7 that it consists in i) an initial segmentation procedure to obtain a first solution and ii) an iterative procedure that tries to modify the previous hypothesis in a way so it will reduce the Description Length. The iteration is stopped when no new hypothesis with a lower DL can be found.

We propose a new strategy to reduce the DL. We use the algorithm introduced in chapters 8 and 9 as an initialisation procedure followed by a DL reduction step. This step relies on an *autonomy*-driven algorithm which we shall now describe.

Given a previous segmentation of the corpus, we define a scoring function for boundary positions. As our initial procedure is based on the maximization of autonomy, any change at any position will result in a lower autonomy of the sequence. Our scoring function evaluates this loss of autonomy whenever a segmentation decision is changed. This can be viewed as being similar to the ordered n-best solutions of our initialisation procedure.

The context of a boundary position is defined as a triple containing:

a position state between two characters, i.e., a boolean set to *true* if the position is a word boundary,

a prefix which is the sequence of characters running from the previous word boundary to the position,

a suffix which is the sequence of characters running from the position to the next word boundary.

When scoring a position, there are two possibilities:

- the position is currently a word boundary: we consider the gain in DL to decide whether the current boundary at that position should be retained or whether we should merge the two surrounding words into just one single word.

- the position is currently not a word boundary. Symmetrically, we test whether or not the DL justify adding a new boundary at that position, splitting one word into its prefix and suffix as two words.

In order to compute the difference in autonomy scores between the current segmentation and the one which is obtained only by performing a merge at one particular position, we simply have to subtract the autonomy of the prefix and suffix and to add the autonomy of the concatenation of the two strings.

Similarly, to evaluate a splitting decision we have to add the autonomy of the prefix and suffix and to subtract the *autonomy* of the concatenation of the two strings.

Note that with this scoring method and this definition of a context as a tuple, all occurrences of a context type will have the same score, and can therefore be grouped together. We can thus evaluate the effect of changing the segmentation decision for a set of identical positions in the corpus in just one step.

Like the lexicon cleaning procedure by Zhikov et al. (2010), we can evaluate the effect of a large number of changes at the same time. But contrary to them, because we process the whole corpus and not the lexicon, we have a broader search space which allows for the creation of large words even if they were previously absent from the lexicon.

A remaining issue is that changing a segmentation decision at a particular position should result in a change of the scores of all the neighbouring positions inside its *prefix* and its *suffix* and require to rebuild the whole agenda, which is a costly operation. To make our algorithm faster, we use a simplified processing that freezes the affected positions and prevents further modification (they are simply removed from the agenda). As the agenda is sorted to test the more promising positions first (in terms of autonomy), this trade-off between exhaustiveness for speed is acceptable. Indeed, it turns out that we reach lower description lengths than (Zhikov et al., 2010).

The details of our minimization of DL algorithm using this scoring method are presented in Figure 10.1. As we shall see, this system can be further improved. We shall therefore refer to it as the *base MDL system*.

10.2. Evaluation of the Base System

10.2.1. Quantitative Results

The results of our base system, without and with our MDL step, are presented in Table 10.1. We also provide results for our re-implementation of the algorithm

Algorithm 10.1.1: ALGORITHM1(*Corpus*)

```
seg ← AUTONOMYMAXIMISATION(Corpus)
DL ← DESCRIPTIONLENGTH(seg)
MinDL ← ∞
Agenda ← SORTBOUNDARIES(Corpus, seg)
while DL < MinDL
do { MinDL ← DL
     for each changes ∈ Agenda
     do { changes ← removeFrozen(changes)
          newDL = SCORE(changes)
          if newDL < MinDL
            then
          do { seg ← ApplyChange(changes)
               FREEZE(changes)
               DL ← newDL
               break
```

Figure 10.1.: **DL minimization**

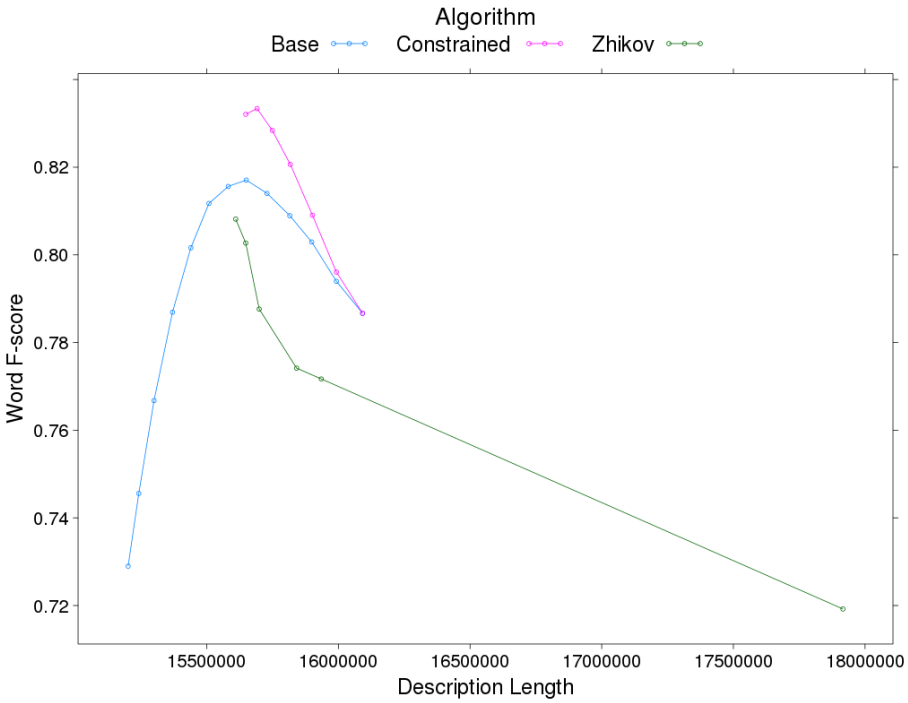


Figure 10.2.: **f-score on words as a function of description length for the three algorithms**

Method	**f-score**	**DL** (Mb)
PKU corpus		
Zhikov *et* al.	**0.808**	15.6
N_μVBE	0.786	16.1
N_μVBE + MDL	0.729	**15.2**
*G*old	*1.0*	*15.0*
City-U corpus		
Zhikov *et* al.	**0.787**	19.8
N_μVBE	0.772	20.3
N_μVBE + MDL	0.749	**19.3**
*G*old	*1.0*	*19.0*
MSR corpus		
Zhikov *et* al.	**0.782**	31.9
N_μVBE	**0.782**	33.0
N_μVBE + MDL	0.693	**31.1**
*G*old	*1.0*	*30.8*
AS Corpus		
Zhikov *et* al. (with their MDL)	**0.762**	67.1
N_μVBE	0.757	68.9
N_μVBE + MDL	0.711	**65.7**
*G*old	*1.0*	*65.3*

Table 10.1.: **Scores on different Corpora for Zhikov et al. (2010) algorithm (without and with their MDL-based improvement step) and for our base system (without MDL and with our base MDL algorithm). Final results are displayed in Table 10.3**

by (Zhikov et al., 2010), without and with their own MDL step, already given in chapter 7. Our initialisation (without our MDL step) obtains very good results; on the MSR corpus, they are even as high as the results of Zhikov *et* al.'s full algorithm, including their MDL step. However, at a first sight, the results we get when using our MDL procedure are disappointing: it may worsen the results of the initialisation step. However, we observe that our MDL step successfully decreases the Description Lengths obtained after the initialisation step, and leads to Description Lengths lower than Zhikov et al. (2010)'s system although with lower f-scores. This tackles the common idea that lower Description Length always yields better segmentation, and

calls for further analysis.

10.2.2. Step-by-step MDL results

In both systems, ours and Zhikov et al. (2010)'s, the MDL algorithm is iterative. We therefore decided to dump intermediary results at each iteration to observe the evolution of the segmentation quality as the DL gets smaller. Figure 10.2.1 shows the resulting f-scores as a function of the DL at different stages, on the PKU corpus (results on other corpora behave similarly). Each iteration of one MDL algorithm or the other reduces the DL, which means that a given curve on this graphic are followed by the corresponding system step after step from right to left. The leftmost dot on each curve corresponds to the point when the corresponding system decides to stop and produce its final output.

This graphic shows that our system produces better segmentation at some point, outperforming Zhikov et al. (2010)'s system. But it doesn't stop at that point and the f-score drops as the DL continue to decrease. This seems to mean that our algorithm, because it explores a larger search space, manages to find segmentations that are optimal as far as DL is concerned, but that do not constitute optimal word-level segmentation.

In order to better understand what is going on, we added a logging functionality to our implementations, so we can check which operations are made when the f-score decreases. We discuss several typical examples thereof.

10.2.3. Error Analysis

A sample of the latest modifications made by our system while the f-score is falling is given in Table 10.2. We show the modifications that are applied to the largest numbers of occurrences. The type of operation is either a merge (suppression of a boundary) or a split (adding a boundary). We provide the prefix and suffix, whether the merge or split is an error or not, as well as English glosses.

The first observation we make is that amongst highly frequent items, our system only performs merges. Splits are indeed performed on a large number of rare types for which both the prefix and the suffix exist in the lexicon. We note that for bigrams, such splits are almost always erroneous.

Merge operations include valid decisions, erroneous decisions producing multi-word expression units (MWE), and erroneous decisions that merge a grammatical word to one of its collocations.

String	Evaluation
的 . 发展 DE - development	error
的 . 问题 DE - question	error
据 . 新华社 According to - Xinhua Agency	error
新华社 . 北京 Xinhua Agency - Peking	error
经济 . 发展 economic - growth	MWE
进行 . 了 conduct - LE (-ed)	error
和 . 发展 AND - development	error
在 . 北京 AT - Peking	error
邓小平 . 理论 Deng Xiaoping - Theories	MWE
领导 . 干部 leading - cadre	MWE
精神 . 文明 spiritual - civilization	MWE
常 . 委会 standing - committee	MWE
反 . 腐败 anti - corruption	correct
节 . 日 holi-day	correct

String	Evaluation
党 . 中央 central committee	correct
金融 . 危机 finance - crisis	MWE
新 . 世纪 new - century	error
副 . 总理 vice - premier	correct
国民 . 经济 national - economy	MWE
马克思 . 主义 Marx - ism	?
北京 . 市 Peking - city	correct
基础 . 上 basis - postposition (=basically)	error
副 . 主席 vice-chairman	correct
经济 . 发展 economic - growth	MWE
结构 . 调整 structural adjustment	MWE
产业 . 化 industrial - ize	correct
现代化 . 建设 modernization - drive	MWE
人 . 大 Acronym for Renmin University	correct

Table 10.2.: **Examples of merges (sorted by number of occurrences)**

Method	**f-score**	**DL** (Mb)
PKU corpus		
Zhikov *et* al. (with their MDL)	0.808	15.6
$N_\mu VBE$ + constrained MDL	**0.826**	15.6
Gold	*1.0*	*15.0*
City-U corpus		
Zhikov *et* al.	0.787	19.8
$N_\mu VBE$ + constrained MDL	**0.801**	19.8
Gold	*1.0*	*19.0*
MSR corpus		
Zhikov *et* al.	0.782	31.9
$N_\mu VBE$ + constrained MDL	**0.809**	32.1
Gold	*1.0*	*30.8*
AS Corpus		
Zhikov *et* al.	0.762	67.1
$N_\mu VBE$ + constrained MDL	**0.795**	67.3
Gold	*1.0*	*65.3*

Table 10.3.: **Final results**

10.3. Description and Evaluation of our Constrained System

Given this error analysis, there are three main types of common mistakes that we would like to avoid:

- merging MWEs such as named entities;
- merging function words with content words when the co-occurrence is frequent;
- splitting bigrams that were correct in the initial segmentation.

If we give up on having a strictly language-independent system and focus on Mandarin Chinese segmentation, these three issues are easy to address with a fairly low amount of human work to add some basic linguistic knowledge about Chinese to the system.

The first issue can be dealt with by limiting the length of a merge's output. A MWE will be larger than a typical Chinese word that very rarely exceeds 3 characters. With the exception of phonetic loans for foreign languages, larger units typically correspond to MWE that are segmented in the various gold corpora.[1] The question whether it is a good thing to do or not will be raised in the next section, but for a higher f-score on word segmentation, leaving them segmented does help.

The second issue can be addressed using a closed list of function words such as aspectual markers and pre/post-positions. As those are a closed list of items, listing all of them is an easily manually tractable task. Here is the list we used in our experiments:
的、了、上、在、下、中、是、有、和、与、和、就、多、于、很、才、跟

As for the third issue, since Chinese is known to favour bigram words, we simply prevent our system to split those.

We implemented these three constraints to restrict the search space for our minimization of the Description Length an re-run the experiments. Results are presented in the next section.

10.3.1. Evaluation and Discussion

The scores obtained by our second system are given in Table 10.3. They show a large improvement over our initial segmentation and outperform previously reported results.

The results presented in this chapter invite for discussion. It is well accepted in the literature that MDL is a good indicator to find better segmentation but our results show it is not always the case. It is possible to reach a lower description length without improving the segmentation score. However, our results also demonstrate that MDL can still be a relevant criterion when its application is constrained using very simple and almost zero-cost linguistic information.

The constraints we use reflect two underlying linguistic phenomena. The first one is related to what would be called "multi-word expressions" (MWE) in other scripts. It is unclear whether it is a limitation of the segmentation system or a problem with the definition of the task. As we discussed in chapter 4, there is a growing interest for MWE in the NLP community. It is not clear whether it is an issue for our system or for the evaluation method.

The second restriction concerns the distinction between content words and

[1] A noticeable exception are the 4-characters idioms (chengyu) but they seem less frequent than 2+2 multiword expressions.

grammatical words. It is not so surprising that open and closed wordclasses show different distributions and deserve specific treatments. From a practical point of view, it is worth noting that MDL is useful for open classes where manual annotation or rule-based processing are costly if even possible. On the other hand, rules are helpful for small closed classes and represent a task that is tractable for human, even when facing the need to process a large variety of sources, genres or topics. This division of labour is acceptable for real-world applications when no training data is available for supervised systems.

Another way to view this issue is to consider a more subtle definition of wordhood than binary segmentation. If we relate these grammatical words to the discussion of chapter 3, we can argue that most of them indeed exhibit only a weak autonomy. Being able to capture this phenomenon could also be considered a good behaviour but would require a more subtle evaluation.

Nevertheless, the bad results of the unconstrained system can also receive the same interpretation as the Bayesian DP model (Goldwater et al., 2009). Both try to optimise a unigram model, considering that the corpus is generated by a random procedure equivalent to drawing tokens from a "bag of word" (the lexicon), each word as an occuring probability $P(w|M_w)$ that depends on the lexicon M_w but not on the context in which it occurs. Higher order models have been proven more accurate on the segmentation task in works on Bayesian inference (see chapter 7). The mathematical resemblance of the two paradigms suggests that we may expect similar findings with context-sensitive models. However such models would not allow for the same optimisations as the simpler unigram model. This is thus a promising yet difficult trail for further improvement. Last but not least, if a higher order model is a desirable feature, Fourtassi et al. (2013) tend to demonstrate that what exactly the appropriate order should be is still unclear.

Of greater concern to us is whether the output of MDL-augmented system is relevant for the linguistics considerations of the first part of our dissertation. The interesting findings on MWE can more straightforwardly achieve by considering the n-best results of our initial segmentation, and the MDL procedure is only a way to select from these. Another interesting aspect in the output of the MDL augmented system is the probabilist lexicon built from the *autonomy* measure that could be used in further inference. The question whether it is worth the computation complexity or not is left for future works.

Chapter 11

Qualitative Evaluation

In the previous chapters, we proposed segmentation procedures based on a simple linguistically sound measure. The good results we obtain on traditional NLP datasets and the standard evaluation methodology tend to demonstrate the validity of our approach. On the other hand, we argue in the first part of this dissertation that the standard evaluation methodology cannot account for all of our concerns.

In this chapter, we propose two other ways to qualitatively judge our outputs. The first one relies on a manually annotated treebank for typing each boundary or non-boundary between each pair of characters with morphosyntactic informations. The second one is to provide a visualisation that includes the n-bests segmentations of our system.

Finally, we provide a method to turn our combinatoric *autonomy* measure into a probabilistic formulation. This will ease it combination with other clues in future works.

11.1. Evaluation of Typed Positions

For this evaluation, we decide to make use of deeper manual annotation available through the Sinica Treebank (Huang et al., 2000).

The Sinica Treebank (hereafter STB) is a corpus made available by the Academia Sinica (Taiwan). Its annotation is based on a "Information Case Grammar" which includes part-of-speech (POS) tagging and phrase-structure analysis where heads are marked, along with the thematic roles of their arguments.

We rely on the head and thematic roles annotation to derive dependency trees which allows us to type positions between each pairs of characters in the corpus. This evaluation is thus only done on positions, similarly to the P_b and R_b scores. The only difference is that we distinguish boundary positions from non-boundaries positions.

For a given type, the correct answer will always be the same (either a boundary or not), therefore we don't use the F-score but a simple accuracy score

$$\text{Accuracy}_{\text{type}} = \frac{\#\text{good answers of this type}}{\#\text{answers of this type}}$$

11.1.1. Typing Positions

The first step to conduct this evaluation is to define how to type positions. There are two main cases: a position is either a boundary in the gold standard or it is not.

If a position is not a boundary, we use the POS tag of the word that contains it. This allows us to make statements such as "*X%* of our decisions are correct inside words with *part-of-speech Y*".

The list of the POS used in the Sinica Treebank are summarised in Table 11.1. We use the simplified tagset.[1] But we keep subclasses for verbs and nouns.

The STB also provides some "morphological features" added to the POS, such as [+NEG] when a verbal compound is at the negative-potential form or [+ASP] for some aspectual markers considered bounded to a verb.

Example (1) is an example of a single word tagged with the feature [+NEG] and examples (2) shows an example of a single word tagged with the feature [+ASP]. Both have a clear internal structure. Their segmentation are questionable: for the same items, we may have conflicting analysis in the Sinica Balanced Corpus used during the bakekoff 2 and in the Sinica Treebank.

(1) • 走 ∘ 不 ∘ 出 •
• walk ? not ? get out •
'to be unable to walk out somewhere'

(2) • 跳 ∘ 了 ∘ 起來 •
• jump ? LE ? upward/begin •

[1] As suggested here :`http://db1x.sinica.edu.tw/kiwi/mkiwi/modern_c_wordtype.html`

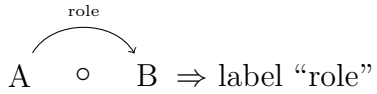
Table 11.1.: **Sinica Simplified Tagset**

Simplified tag	Original tags (subclasses)	name in MSC	translation
A	A	（非謂）形容詞	non predicative adjectives
ADV	D, Da, Dfa, Dfb, Dk	副詞	adverbs
ASP	Di	時態標記	aspectual markers
C	Caa, Cbb	連接詞	conjunctions
DET	Nep, Neqa, Nes, Neu	定詞	determiners
FW	FW	外文標記	foreign script
M	Nf	量詞	measure words
N	Na, Nb, Nc, Ncd, Nd, Nh	名詞	nouns
P	P	介詞	preposition
POST	Cab, Cba, Neqb, Ng	後置詞	postposition
T	DE, I, T	語助詞	auxiliary word
Vi	VA, VB, VH, VI	不及物動詞	intransitive verbs
Vt	SHI, VAC, VC, VCL, VD, VE, VF, VG, VHC, VJ, VK, VL, V2	及物動詞	transitive verbs

'starting to jump around'

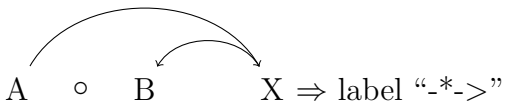
If a position is a boundary, we rely on the thematic relation between the two words before and after the evaluated position to define its type. We can distinguish three subcases:

- If one of the two words is the direct governor of the other (i.e., the former assign a thematic role to the later), we use the thematic role of the dependent to label the position

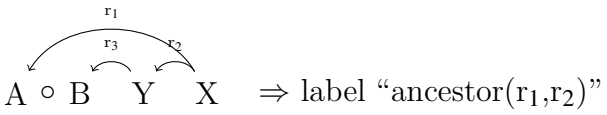


⇒ label "role"

- If one of the two words is an ancestor of the other in the dependency tree (but not its direct governor), we use an "indirect dependency" label to type the position. We use the label "-*->" (resp. "<-*-") if the word higher in the dependency tree is the one on the left (resp. right) of the position.



- If neither of the two previous cases apply, we find the common ancestor of the two words in the dependency tree and use the pair made of the first thematic role on the path from the common ancestor to each word to label the position.



From the set of possible thematic roles used in the Sinica Treebank, we provide the translation of the roles that will appear in our results (we translate the guideline):

addition labels additions and extensions.
(表附加)

agent labels the initiator of an event, actant of an action verb.
(表事件中的肇始者,動作動詞的行動者)

apposition labels the apposed and coreferent objects.
(表物體的同位語,即指涉相同的物體。)

aspect labels the aspect marker of a verb.
(表動作的時貌。)

causer labels the initiator of an event. Only if it is not the agent.
(表事件的肇始者,但肇始者並未主動促使事件發生。)

complement labels afterthought, complementary information
(表補充說明,進一步補充前一事件內容。)

concession labels concessive conjunction. (表讓步語氣的連接。)

condition labels the marker that introduces a condition.
(表條件語氣的句子。)

contrast labels the marker that introduces a turn in the argumentation.
(表轉折語氣。)

degree labels the marker of the degree of a state.
(表狀態的程度。)

deixis labels the deictic component of an action.
(表動作附加的指示成分。)

deontics labels the adverb introducing the attitude of the speaker regarding an event.
(表說話者對事件是否成真的態度,標示於此類型的法相副詞。)

DUMMY, DUMMY1, DUMMY2 For undecided cases and for the parts of a coordination.
(表未定的角色,需要靠其上位詞組的中心語才能決定。若是在並列結構中則又有 DUMMY1 與 DUMMY2 兩個角色以區分前面部分和後面部分。)

duration labels the duration of an event.
(表事件持續的時間長度。)

epistemics marks the degree of certainty of the speaker regarding the event.
(表說話者對事件是否為真的猜測,標示於此類型的法相副詞。)

evaluation marks an evaluative modality
(表評價的語氣成分。)

experiencer marks an agent affected by the emotion or perception described, subjects of emotion verbs. (表感受所敘述的情緒感知狀況的主事者,為心靈類述語的主 語。)

frequency marks the frequency of an event.
(表事件的頻率。)

goal marks the target affected by a verb, or the targeted object of an emotion verb, or the receiver or endpoint of transmission events.
(表動作影響的對象,或者為心靈動作的受事對象,在有物件轉移的事 件中則是個接受者或終點。)

hypothesis (表假設的語氣。)

location marks the place where the event takes place
(表事件發生的地點。)

manner (表主語的動作方式)

negation（表否定）

nominal marks nominalisations, used to annotate the 的 (de) in nominal phrases headed by a verb.
（表名物化結構，用來標示中心語為名物化動詞的名詞短語中的「的」）

particle marks the sentence final particle(表句尾説話者的語氣)

possessor marks the genitive: including element, creator, owner and all other cases corresponding to genitive.
（表物體的領屬者，包含成員、創造者、擁有者和整體等皆為領屬者。）

predication marks a modification, a relative clause attached to a noun.
（表修飾物體的相關事件，為名詞的關係子句。）

property high level or coarse thematic role, marks properties, qualities as well as temporal and spacial informations.（表物體的特色和性質，也包含物體相關的時空訊息，是一個較上位而粗略的語意角色。）

purpose（表目的）

quantifier marks the determiners that quantify nominals.
（表名詞的數量修飾語，為數量定詞、定量詞等等。）

quantity（表事物的數量。）

range mark the belonging to a category or the scope of a result. Principal role assigned by classification verbs and in comparative sentences.
（表分類的範疇或結果的幅度。為分類動詞及比較句的主要語意角色）

reason（表事件的原因）

result（表事件的結果）

theme marks the target of stative or classifying predicates, the agent described as existing or moving in dynamic events as well as patient that are created by the event.
（表靜態及分類述詞敘述的對象或動態事件中描述存在或位移的主事者，以及因事件動作造成物體的狀態從無到有的受事者，皆使用這個語意角色。）

time（表事件發生的時間。）

topic（表事件所論述的主題。）

11.1.2. Experiment Results

The Sinica Treebank is relatively small compared to the corpora we used so far (it consists in only 361,834 tokens). As the sentences analysed in the Treebank were selected from the Sinica Corpus, the same corpus from which the AS corpus from the bakeoff is sampled, we decided to train our system on the union of the STB and the AS corpus. The evaluation is done on the STB part only.

We first use our basic system with preprocessing but without MDL. We only show the 200 more frequent types in the corpus to focus on representative data and limit the number of lines. Results are reported in Table 11.2 and 11.4.

They demonstrate that the quality of the segmentation greatly varies among different types in a way that seems unrelated to type frequencies.

It shows that the boundaries corresponding to long-distant dependencies, that is the "common ancestor" cases labelled ancestor(X,Y), the "-*->" and "<-*-" cases are almost always correctly addressed. This is in line with Tesnière's idea that the cuts in a sentence are not all of an equal depth and that our system performs better on deeper cuts. It seems very likely that boundaries between two remotely connected words correspond to the "deep nicks" of Tesnière.

On the other hand, querying the corpus for cases on which our system performs badly leads to closed classes of frequent items, such as demonstratives or adverbs. In a few cases it may correspond to the "common ancestor" type, for example when the missed boundary is between two adverbs that both depends on a same verb. This is also consistent with the idea that more shallow cuts should be harder to detect. They are also more likely to become formulas and to be stored holistically in the mental lexicon. Last but not least, these items typically exhibit weak autonomy as discussed in chapter 3, some could even be described as indissociable. Although they are highly frequent types, our bad performances on negation and aspect markers could have been expected and are to some extent consistent with our theoretical position and what we try to model. Typically, aspect markers could perfectly be described as verb suffixes. If we decide that we have to segment them as independent units, this can not be done solely on the ground of their *autonomy*.

We already know that our system without MDL exhibits a higher R_b and a lower P_b, we can thus expect more mistakes on the position inside words.

Noteworthily, parts-of-speech on which our system performs the worse have a

clear and regular morphological structure. For example, all those which include a morphological feature in square brackets are "morphologically" derived, that includes potential forms of verbs discussed in chapter 3. The verbs forms tagged VG also exhibit interesting properties as they have a clear Verb+Preposition internal structure as in 稱為 chēngwéi 'to be called-as' but accept the aspect marker 了 le after the second character (稱為了 and not after the first one *稱了為).

We ran the same evaluation on the output of our system with MDL, samples of the results are provided in tables 11.3 and 11.5. Full tables are available in the appendix A.

It shows that the MDL tends to exacerbate the behaviour of the system, erroneously merging more items which could not be kept apart by simple rules and correctly segmenting more long distance dependencies types. On the other hand, it seems to inhibit the tendency of our base system to over-segment. However given the nature of our future work, over-segmentation may not be an important issue and we may prefer the base system which has a clearer formulation and a more predictable behaviour.

11.2. Visualisation of the *n*-best Solutions

The first part of this dissertation underlined the fact that the *autonomy* is only one aspect of the wordhood. We do not hope to achieve a perfect result with this single clue.

In fact we have not reached a convincing computable definition of the "perfect result" yet. We only argued that a first segmentation based on the *autonomy* is required to further analyse word candidates in terms of wordclass memberships. Previous evaluations are not to be considered as a goal in itself but only as evidences that our first steps are made in a sensible direction.

To go further, our base segmentation system does not have to yield a unique solution. We only choose the solution that maximises the *autonomy* overall for the sake of a quantitative evaluation that can be compared with previous works. In our future work on wordclasses, we may consider working on the *n*-best solutions, allowing some ambiguities.

This solution is uneasy to evaluate and compare to other systems but we can provide a visualisation tool. It will not lead to an objective evaluation but will enable us to see how the system would perform if some flexibility were allowed.

As we don't need manually segmented reference in this case, we took this opportunity to experiment on online spontaneous writing, a genre which is still challenging for supervised systems.

11.2.1. The Data

We obtained the recording of months of discussion on the public IRC chatroom of the 零時 (língshí) 政府 (zhèngfǔ) (g0v.tw community[2], a very active group of OpenData enthusiasts in Taiwan).

The content of the discussion is expected to include jargon specific to this social group and to topics related to open data that would be absent from dictionaries and manually segmented corpora from the 90's. On the other hand, we expect the discussion topics to be recurrent. This redundancy should compensate the relatively small size of the corpus.

[2]See http://g0v.tw/about.html,

Accuracy	Type	# occurrences
0.235	ancestor(epistemics,quantity)	183
0.361	ancestor(deontics,quantity)	266
0.383	ancestor(epistemics,evaluation)	324
0.433	ancestor(epistemics,epistemics)	127
0.443	ancestor(negation,evaluation)	553
0.446	ancestor(epistemics,time)	368
0.533	negation	2229
0.549	ancestor(deontics,epistemics)	133
0.572	ancestor(deontics,evaluation)	437
0.591	ancestor(deontics,time)	523
0.636	aspect	3544
0.642	quantity	2674
0.650	ancestor(quantifier,quantifier)	314
0.653	ancestor(negation,quantity)	121
0.680	degree	3199
0.692	evaluation	3282
0.723	ancestor(degree,evaluation)	173
(...)	(...)	(...)
0.966	ancestor(property,quantifier)	1917
0.966	-*->	32239
0.967	ancestor(quantifier,property)	967
0.967	ancestor(time,theme)	1881
(...)	(...)	(...)
0.985	ancestor(predication,predication)	132
0.985	frequency	134
0.985	ancestor(theme,contrast)	405
0.986	ancestor(time,goal)	209
0.987	ancestor(quantifier,predication)	156
0.988	ancestor(dummy,dummy)	502
0.989	ancestor(range,range)	1226
0.991	ancestor(range,aspect)	323
0.991	ancestor(theme,location)	116
0.992	ancestor(time,topic)	121
0.993	ancestor(evaluation,goal)	141
0.993	ancestor(nominal,goal)	149
0.993	ancestor(theme,result)	152
0.995	ancestor(agent,agent)	188
0.995	ancestor(evaluation,theme)	1422
0.995	ancestor(property,apposition)	212
1.000	ancestor(agent,contrast)	241
1.000	ancestor(agent,result)	132
1.000	ancestor(theme,topic)	145

Table 11.2.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, with our basic system (sample from Appendix A)**

Accuracy	Type	# occurrences
0.213	ancestor(epistemics,quantity)	183
0.299	ancestor(epistemics,epistemics)	127
0.327	ancestor(deontics,quantity)	266
0.330	ancestor(epistemics,evaluation)	324
0.389	ancestor(epistemics,time)	368
0.421	ancestor(negation,evaluation)	553
0.466	ancestor(deontics,epistemics)	133
0.490	negation	2229
0.540	aspect	3544
0.545	ancestor(deontics,evaluation)	437
0.556	ancestor(deontics,time)	523
0.595	ancestor(negation,quantity)	121
0.600	degree	3199
0.611	quantity	2674
0.624	ancestor(quantifier,quantifier)	314
0.640	ancestor(agent,evaluation)	178
0.673	evaluation	3282
(...)	(...)	(...)
0.952	<-*-	15116
0.953	ancestor(theme,contrast)	405
0.954	-*->	32239
0.955	ancestor(degree,theme)	733
0.957	ancestor(complement,complement)	257
0.961	dummy1	4592
0.962	causer	130
0.963	ancestor(quantifier,property)	967
0.963	ancestor(theme,aspect)	703
0.963	ancestor(time,theme)	1881
(...)	(...)	(...)
0.980	ancestor(property,nominal)	152
0.982	ancestor(manner,theme)	453
0.983	ancestor(range,range)	1226
0.985	ancestor(predication,predication)	132
0.985	frequency	134
0.986	ancestor(time,goal)	209
0.986	ancestor(dummy,dummy)	502
0.987	ancestor(quantifier,predication)	156
0.988	ancestor(range,aspect)	323
0.991	ancestor(theme,location)	116
0.992	ancestor(time,topic)	121
0.993	ancestor(nominal,goal)	149

Table 11.3.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, with our basic system + MDL (sample from Appendix A)**

Accuracy	type	# occurrences
0.547	VC[+NEG]	243
0.591	VA[+ASP]	159
0.668	Ndaaa	407
0.674	Ndc	132
0.707	VB	1019
0.752	Ndabf	359
0.754	Ndabc	553
0.777	Naa	2104
0.787	VA	8038
0.792	Naea	698
0.796	A	1798
0.798	Nba	15815
0.801	VD	603
0.804	Ncb	9675
0.809	ADV	26623
0.810	VG	1920
0.815	Nab	29195
0.840	DET	5663
0.844	Ndabe	965
0.847	Nad	16472
0.850	Ndaab	214
0.852	VC	15675
0.854	VH	19639
0.862	VJ	4275
0.865	C	850
0.867	VI	498
0.872	Ndabb	257
0.873	Nce	394
0.876	Nca	12247
0.887	Ncdb	1093
0.887	ALL	595561
0.887	Ndabd	1519
0.889	Naeb	3037
0.895	Ndaad	1602
0.898	Ncc	1088
0.903	Nv	3999
0.910	Nac	16547
0.919	Ndda	431
0.919	P	954
0.921	Ndaba	392
0.921	Cbcb	430
0.927	VF	750
0.946	VE	4754
0.949	VK	3221
0.958	VL	872
0.968	POST	2405
0.976	Cbca	1205
0.979	Nddc	1176
0.985	P0	342
0.987	Nhab	1050
0.993	ASP	147
0.998	Nhaa	1867
1.000	Nddb	325

Table 11.4.: **Typed positions evaluation on non-boundaries, basic system**

Accuracy	type	# occurrences
0.617	VC[+NEG]	243
0.654	VA[+ASP]	159
0.668	Ndaaa	407
0.758	Ndc	132
0.792	Ndabc	553
0.793	VB	1019
0.838	Ndabf	359
0.850	ADV	26623
0.859	Nba	15815
0.863	A	1798
0.874	Ndabe	965
0.878	VA	8038
0.883	DET	5663
0.894	VD	603
0.894	VG	1920
0.896	Ncb	9675
0.898	Ndaad	1602
0.899	Ndabb	257
0.899	ALL	595561
0.904	Naa	2104
0.909	Nab	29195
0.919	VH	19639
0.919	Nca	12247
0.921	VC	15675
0.922	VJ	4275
0.928	Nad	16472
0.943	Naea	698
0.943	Ncdb	1093
0.944	Ndaab	214
0.944	Ndabd	1519
0.944	Cbcb	430
0.946	VI	498
0.948	C	850
0.949	Nce	394
0.951	Ndda	431
0.952	Ndaba	392
0.952	Ncc	1088
0.954	Nv	3999
0.962	Naeb	3037
0.963	Nac	16547
0.972	VF	750
0.973	P	954
0.974	VE	4754
0.980	VK	3221
0.984	VL	872
0.990	POST	2405
0.990	Nddc	1176
0.991	P0	342
0.993	ASP	147
0.994	Cbca	1205
0.994	Nhab	1050
0.999	Nhaa	1867
1.000	Nddb	325

Table 11.5.: **Typed positions evaluation on non-boundaries, basic system +MDL**

11.2.2. Visualisation Tool

We choose to use *What's wrong with my NLP*[3] to visualise the output of our system. It is designed to visualize various kind of annotation on corpora, especially properties over a (contiguous) span of tokens and relations between tokens.

We display characters as tokens and use span properties to display the base segmentation and autonomy values for subsequences of the segmented words below the sequence of tokens. We use the relation links to show possible units larger than our base segmentation. We simply draw a link from the first to the last character of autonomous sequences larger than our base segmentation units.

In other words, if the 1-best output of our system is right, spans below the texts can be considered as "morphology" and links over it as "syntax". Note that the links are not to be taken as tentative dependency relations, they are to be seen as covering potential phrases or MWE.

We also use relation links to enrich the visualisation with dictionary lookup. This allows us to easily see what we may be missing or what we capture that a dictionary could miss.

We use the free and collaborative CEDict[4] Chinese-English Dictionnary. We label the links with the pinyin romanisation and a click on the link shows the definition. We only show the words of length 2 or more as the dictionary include etymological meaning for non-autonomous characters without a clear distinction from autonomous forms.

This visualisation is illustrated on Figure 11.1. The sentence in the illustration is glosed in (3).

(3)

奇怪	為	什麼	空白	字元	會	變	方框
qíguaì	*wèi*	*shénme*	*kòngbái*	*zìyuán*	*huì*	*biàn*	*fāngkuāng*
Weird	for	what	blank	character	may	change	square

'How weird, why do the blank characters may become squares.'

In the figure, boxes without numbers indicate the choice of our base algorithm that leads to an overall maximisation of the *autonomy*. Links with Latin characters indicate available dictionary entries. All other information comes with the *autonomy*

[3] `https://code.google.com/p/whatswrong/`
[4] `http://cc-cedict.org/wiki/`

value of the corresponding string.

We see that a mistake made by our system concerns the word 字元 zìyuán 'character' which is over-segmented. Both characters of this word may appear in isolation. 字元 zìyuán is specific to characters in computer systems. 字 zì is a generic word for characters and 元 yuán a character used to denote various units, especially amounts of money. This leads our system to an over-segmentation, but we see that it also considers 字元 zìyuán as a possible solution with a higher autonomy value than 元 yuán alone. In this case, the very high *autonomy* value for 字 zì is responsible for the mistake.

The other mistake is to consider 變方框 biànfāngkuāng 'to become square' as a single word. Here also, the second-best solution would have been correct.

As expected, our system tends to over-segmentation.

Let us now consider the links on the upper side of the figure. If we consider those of higher *autonmy*, we find interesting segments. Some are included in the dictionary, other could have been:

- 為什麼 wèishénme 'Why' (*lit. for-what* which is indeed impossible to split).
- 奇怪為什麼 qíguàiwèishénme 'How comes ... ?' (*lit. Weird-for-what ?* could eventually become a *formula*.
- 空白字 kòngbáizì 'blank character'
- 空白字元 kòngbáizìyuán 'blank character' (computer domain)

11.2.3. Algorithms

To obtain the representation of the data we just described, we used three simple algorithms.

Dictionnary lookup We simply look for all the substring of length 2+ in the CEDict

Over-segmentation (spans) We compute and display the *autonomy* of all substrings of all the words in 1-best output of the system.

Under-segmentation (links) We modify our basic segmentation procedure to memorise the *n*-best segmentations (using a beam-search). We also discard word candidates that have a negative *autonomy*.

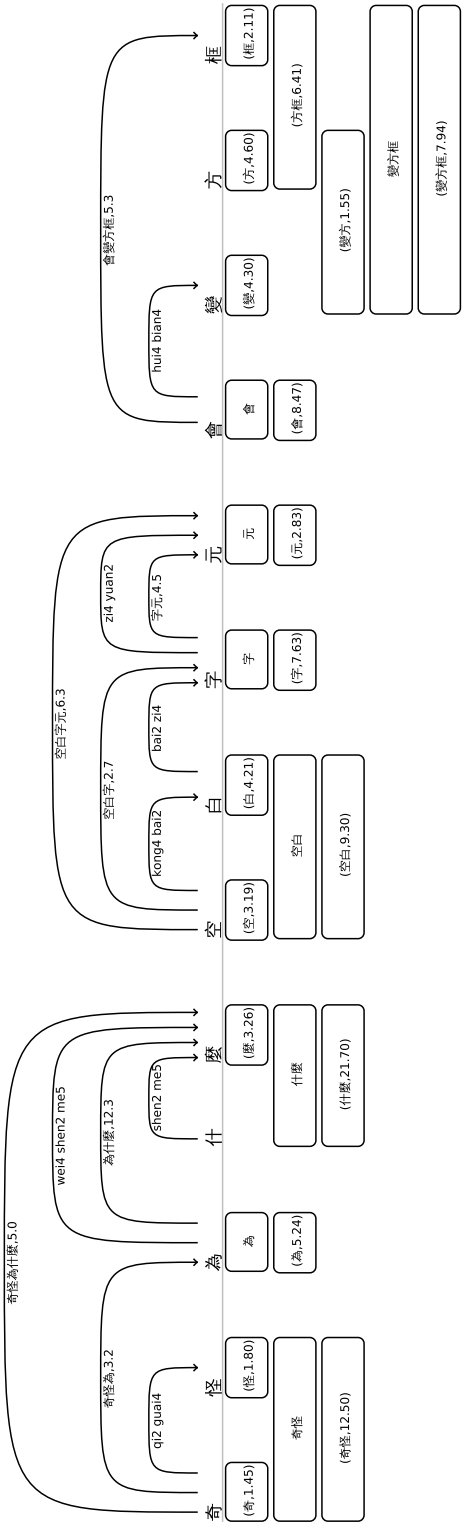


Figure 11.1.: **An example of ambiguous output visualised with *What's Wrong with my NLP***

11.2.4. Discussion

The observed output is very promising for our longer term purpose to pursue with distributional analysis of the candidate wordforms. Although we don't reach a unique correct solution, modifying the system to come up with an ambiguous output that includes alternative solutions allows us to capture almost all the desirable candidates without introducing too much noise in our candidate list.

For example, it is desirable to conclude that 字元 (zìyuán) is a single unit and reject the two words 字 (zì) • 元 (yuán) analysis, but it is sensible to at least consider it on the ground of the *autonomy of its parts*. From a linguistic point of view, rejecting this solution requires further considerations that involve distributional analysis. In a symmetric way, 為 (wèi) • 什麼 (shénme) was a correct answer but conserving 為什麼 (wèishénme) for further analysis and delaying the decision does not sound totally absurd. Such cases are numerous in the corpus.

More problematic cases involve closed sets of frequent items such as adverbs, determiners or pronouns. They are very likely to be erroneously attached to other items if we loosen segmentation criteria. The issue is that there is no obvious way to distinguish such sequences from desirable merges or formula made of lexical items based solely on the *autonomy* values. This is likely to become an important question for future work. On the other hand, this may turn out to be useful to define the so-called *substitution frames* discussed in chapter 3 as they may be good indicators of the word classes of the words they are merged with.

11.3. Bootstrapping a Probabilistic Lexicon

In the first part of this dissertation, we explained that the *autonomy* measure developed in this part is only one of the two indicators for wordhood together with word classes. Segmentation algorithm based solely on *autonomy* are not to be taken as the ultimate goal, but only as a first step required to start a study on word classes. The decision about the wordhood of a sequence of characters should be taken according to a combination of both types of clues. A study on word classes and its combination with the insights from autonomy will more likely be conducted on the list of potentially autonomous forms (a tentative lexicon) rather than on a segmented corpus.

After the different experiments made to assess and improve our proposal for an

autonomy measure, we can now review the information that was induced for each sequence of character in our raw corpus:

1. the form of the (candidate) word w;
2. its *autonomy score* $a(w)$;
3. the right and left components of the *autonomy score* ($NVBE_{\rightarrow}(w)$ and $NVBE_{\leftarrow}(w)$), these two pieces of information are still useful to keep along with $a(w)$ to study affixes;
4. its occurrences count as a n-gram (segmented or not);
5. its occurrences count as a segmented word when using the maximisation of *autonomy* algorithm;
6. $P(w|M_w)$, the probability to draw this word if we randomly draw a word from the corpus. This is a value used to compute the Description Length, however we can observe it before and after the minimisation of the DL.

Probabilistic modelling will provide a sound and robust framework to work with when combining our definition of *autonomy* with other insights, such as word class memberships which can also be expected to be formulated as a probabilities. It thus seems interesting to present a simple way to turn our present results into a probabilistic formulation.

Under our current definition, $a(w)$ can be any real number. Let us now show how we can derive the "probability that w is a wordform given a certain score $x = a(w)$" from our results.

By the "probability that w is a wordform", we mean how likely it is to be segmented by our algorithm. We call $isWord(w)$ a boolean function that is True if w appears as a word in our unsupervisedly segmented corpus and False otherwise. We are looking for the probability density function

$$pdf(isWord(w)|a(w) = x),$$

We can derive a simpler way to compute this value by applying the Bayes' rule:

$$pdf(isWord(w)|a(w) = x) = \frac{P(a(w) = x|isWord(w))\,P(isWord(w))}{P(a(w) = x)}.$$

In this formulation, $pdf(a(w) = x|IsWord(w))$ and $pdf(a(w) = x)$ correspond to the probability density function we showed in chapter 8, except that we use our unsupervised segmentation to distinguish between words and non-words rather than the manual segmentation. $P(isWord(w))$ can also be estimated from the counts made on our unsupervised segmentation of the corpus.

It could also be possible to condition these probabilities on the length of w.

This enables us to associate a single probability to each observed wordform to build a tentative lexicon without having to decide explicitly a specific segmentation of the corpus.

Chapter 12

Conclusion and Perspectives

To conclude this dissertation, we will firstly give a summary of the results obtained so far.

Secondly, we will present some preliminary experiments conducted in order to test the extendibility of our system to other languages and argue in favour of its genericity. Thirdly we provide some insights into possible research directions that could improve or make use of our present work.

Finally, we shall discuss our findings from a more general point of view.

12.1. Summary

In this work, we discussed the notion of wordhood with a specific focus on Modern Standard Chinese. It appears to us that distinguishing syntactic units and semantic units is necessary for a proper treatment of minimal units for linguistic analysis. To identify the minimal units of syntax, a first segmentation is needed to bootstrap further analysis. From the literature, we retained the Harrissian hypothesis as the most linguistically sound starting point for unsupervised word segmentation. We refined its adaptation for data-driven linguistic analysis, and proposed a novel segmentation algorithm that greatly improved on previous comparable works based on the same initial hypothesis.

The use of Minimum Description Length in chapter 10 can be seen as a re-ranking procedure for the n-best outputs of our system. Our experiments with MDL have shown mixed results. On the one hand we achieved *state-of-the-art* performances with respect to traditional evaluation methodology, but on the other hand doing so

involved *ad hoc*, language-specific adaptations. Although we reached our highest scores with the help of MDL, our experiments actually questioned its relevance for the task. We explain the poor performance of a vanilla-MDL by the fact that it tries to fit a too simplistic model to the task. This is in line with previous findings from Bayesian inference.

We also discussed the limitations of traditional evaluation methods for supervised CWS when applied to unsupervised systems and proposed finer-grain metrics as well as a visualisation tool for a qualitative analysis of the results. As the next step after this work is to find distributional classes, and unlike the traditional Chinese Word Segmentation task in NLP, it is not necessary for us to aim at a non-ambiguous binary segmentation. We can allow for an ambiguous output by taking the n-best outputs of our system. We can also focus on the lexicon rather than on the corpus and we have seen how we can turn our real-valued *autonomy* score into a probability of being autonomous that implicitly accounts for the segmentation of the corpus. This will enable us to use probabilistic models in our future work when combining clues from different sources.

12.2. Extendibility to Other Languages

Although our experiments focus on MSC written with Chinese characters, we believe that the general ideas underlying our system are not restricted to this experimental setup. It is indeed directly usable in other contexts.

We conducted some further experiments with other languages and scripts. Although they are not as thoroughgoing as our experiments on MSC, they show that the proposed system exhibits desirable properties for our future work.

12.2.1. Taiwanese Hokkien

Hokkien is a language spoken in the South of China, amongst Chinese migrant communities and in Taiwan. For a comprehensive presentation of the language, we invite the reader to refer to (Klöter, 2005). We will restrict the presentation to a few properties relevant for our subject. Taiwanese Hokkien is genetically related to Mandarin Chinese (the latter being the basis of MSC). It can be written with Chinese

characters and thus provides a challenge similar to the segmentation of MSC but with a different language. In this experiment, we face roughly speaking the same character set and many linguistic typological properties.

A very interesting fact about Taiwanese Hokkien for our subject matter is that it can be written using Chinese characters, the Latin alphabet or a mixed script that uses both. Texts are widely available in various scripts[1] and this could pave the way to many research possibilities.

For now, we simply use this property to retrieve a second level of *orthographic segmentation*. Unlike romanised Vietnamese which kept the same *orthographic boundaries* (on the syllables) when romanisation replaced Chinese characters, romanised Taiwanese uses hyphens to group multiple syllables together and spaces to separate "words". This experiment is run on the the "Digital Archive for Written Taiwanese" (Iûnn, 2007, DAWT) We work on a corpus which provides an alignment of the Chinese script and its romanised form.[2] This allows us to derive a "gold" segmentation of the Chinese script from the alignment in a way similar to (Iûnn et al., 2009) and to run the same experiments as for MSC.

The DAWT corpus is divided into genres and eras. We provide the results for each subpart. Respective sizes of the data we use in each sub-corpus are reported in Table 12.1.

We have not reached the same level of preprocessing as for MSC yet, so the results presented here are therefore to be compared to the results of chapter 8. They are presented in Table 12.2. We can see that our system obtains scores similar to those obtained on MSC.

From this result, it follows that our system is general enough to deal with a variety of Sinitic Languages and not only MSC. It would require designing proper local grammars specific to each language, but as we have seen it can greatly improve the performances of the system with only a small amount of manual work.

This is especially interesting as Sinitic Languages other than MSC can be considered as under-resourced languages. Less effort and money are invested in the

[1] Unlike the pinyin transcription used to give the pronounciation of Chinese Characters in MSC which has only very seldomly been used in publication and during a very short timespan, the romanisation of Taiwanese is an actual orthography which was and is still in use in publications.

[2] We are extremely grateful to Iûnn Úngiân who kindly provided the data to conduct this experiment

Sub-corpus	# tokens	# words
Qing era	0.16M	0.12M
Japanese era	0.84M	0.64M
Post-war era	1.72M	1.3M
Total	2.72M	2.06M

Table 12.1.: **Size of the Taiwanese data used**

development of NLP tools and resources for these languages. Unsupervised methods to process and help the description of these languages are thus even more interesting than they are for MSC.

12.2.2. Segmentation from Phonemic Transcriptions and Alphabetic Scripts

Chinese script essentially transcribes syllables. More experiments are therefore required to see whether our system can be extended to process phoneme-based scripts or phonemic transcriptions.

The situation is formally very similar: the input is still a sequence of discrete symbols. But respective sizes of the sets of input symbols and lengths of the targeted segmentation units are quite different. In the case of MSC, input corresponds to a few thousand symbols, whereas it amounts to only a few dozens in the case of phonemic transcriptions. The length of the targeted units varies greatly from one language to another.

Although written Thai and French are not phonemic transcriptions, segmentation unit lengths and sizes of the input symbol sets are of the same order as in English phonemic transcription. This is why we discuss these cases altogether.

Phonemic Transcription: the Bernstein-Ratner Corpus

We presented the use of the Bernstein-Ratner corpus (Bernstein-Ratner, 1987) in computational psycholinguistic research in section 7.2.2, where we stressed the

Method	F_w	F_b	R_b	P_b
Taiwanese Hokkien (all data)				
$N_\mu VBE$, sum, no MDL	0.744	0.911	0.895	0.926
Taiwanese Hokkien (Qing era)				
$N_\mu VBE$, sum, no MDL	0.766	0.917	0.952	0.885
$N_z VBE$, sum, no MDL	0.724	0.899	0.933	0.867
Taiwanese Hokkien (Japanese era)				
$N_\mu VBE$, sum, no MDL	0.774	0.921	0.944	0.900
$N_z VBE$, sum, no MDL	0.747	0.909	0.942	0.879
Taiwanese Hokkien (Post-war era)				
$N_\mu VBE$, sum, no MDL	0.749	0.915	0.954	0.879
$N_z VBE$, sum, no MDL	0.730	0.905	0.944	0.870

Table 12.2.: **Segmentation of Taiwanese Hokkien written with Chinese characters, evaluated against the orthographic segmentation of its romanisation.**

differences between psycholinguistic modeling and our work in terms of input data, objectives and constraints. We can nevertheless test our system on the same data.

No modification is needed to do so. The BR Corpus is substantially smaller than all the other corpora we used in this dissertation. But it is also less diverse and the average length of each utterance is typically smaller that the sentences in MSC.

Our results are presented in Table 12.3.

Method	F_w	F_b	R_b	P_b
Bernstein-Ratner Corpus				
DP (Goldwater et al., 2009, unigram)	0.538	0.743	0.622	0.924
HDP (Goldwater et al., 2009, bigram)	0.752	0.852	0.808	0.903
(Zhikov et al., 2010, with n=2)	**0.760**	**0.926**	0.936	0.916
$N_\mu VBE$, sum, without MDL	0.696	0.909	**0.942**	0.879
$N_\mu VBE$, sum, with MDL	0.580	0.881	0.840	**0.927**

Table 12.3.: **Segmentation of phonemised Child-Directed Speech from the BR Corpus**

On this corpus, our system obtains its lowest scores. We can nevertheless note

two points of interest: i) the base system obtains a high boundary recall rate, which means that it finds more correct boundaries than other systems but tends to over-segment and ii) DP (Goldwater et al., 2009, unigram model) and our MDL have similar performances, exhibiting low F-scores but high boundary precision. This is in line with the fact that both DP and our MDL try to find an optimal solution for (inadapted) probabilistic unigram models.

The MDL procedure from (Zhikov et al., 2010) is unable to merge longer sequences by-design. This prevents it from exploring the solutions found by DP and our MDL.

"Alphabetic" Writings: Thai and French

The Thai language and script belong to the set of languages written without word boundary punctuation and thus requires a segmentation step. Shared tasks were thus organised in the NLP community on Thai Word Segmentation for which manually segmented evaluation data was made available. This enables us to perform a quantitative evaluation on Thai.

Our results show the same properties as those on the BR corpus. Given our lack of knowledge about the Thai language, it is difficult to elaborate on them.

In order to be able to interpret the output[3], we also tried our system on extracts from the French Wikipedia from which we removed the white-spaces. Samples of the output and the induced lexicon are given in the appendix B.

There is no way to properly evaluate this output quantitatively and the segmentation is far from being a perfect match to orthographic spaces. But given the fact that we aim only to recognise autonomous sequences, the output seems reasonable. In most of the cases where the system is wrong, the segmented sequence corresponds to another autonomous form. To be able to perform contextual disambiguation needed to correct those cases, we would need deeper syntactic insights.

Although the scores we obtained in other traditional NLP tasks are not appealing, our analysis of the results allows us to remain confident that our system is efficient for our purpose.

[3]The comments about French are not expected to describe the situation for Thai

12.3. Going further

At that point, we have a new method to induce either i) a segmentation of a corpus into units that may commute or ii) a list of forms ordered by their estimated *autonomy*.

We shall now consider possible future directions.

12.3.1. Semi-supervised Learning

In some cases, for example if a syntactic parser requires an unambiguous tokenisation as its input, we may need a single segmentation that follows specific guidelines.

But if we do not have a manually segmented corpus of the proper domain or genre at hand, developing one that would be large enough to train a supervised segmentation system can be costly and time-consuming.

It should be possible to use the ideas and measures presented in this dissertation to reduce the size of the manually segmented data required for training (provided that a larger unannotated corpus is available for the unsupervised component).

We see two possibilities in this direction. The first one is to estimate the *autonomy* score of the sequences in the corpus based on the larger corpus and to rely on these values to train a supervised system. The second one is to segment a corpus in an unsupervised way and to use the manually segmented data to train a system that would only have to correct the output of the unsupervised segmenter.

We explored the first solution using Conditional Random Fields (CRF) for the supervised machine learning algorithm and our measure of *autonomy* as a feature for training. In doing so, we achieve performances which rival other supervised systems when trained on large datasets but we were not able to maintain this high performance when the size of the training data was reduced.

The second solution seems more likely to succeed. Using the ambiguous output from section 11.2, we could now adapt the TBL system (Brill, 1995) as proposed by Gao et al. (2005) which we mentioned in section 7.1.3. The scores we obtain with our unsupervised segmentation are close to the scores reported by Gao et al. (2005) before the adaptation to a specific guideline, and our ambiguous output can be compared to their trees describing the internal structures of words. This could be a better solution than the CRF-based option we tried without success.

12.3.2. Multi-Word Expressions

As we mentioned in chapter 4, MWE is an active area of research in NLP. With the use of manually crafted external resources that describe MWE, it is possible to improve NLP systems performing various tasks such as syntactic parsing (Constant et al., 2011; Green et al., 2011). Our system could replace or improve such manually designed lexicons in various settings.

To do so, we could simply treat *orthographic words* in Latin scripts as input symbols and run our system *as it is*. The resulting ordered multi-word lexicon seems reasonable but we have not conducted a proper evaluation yet. We believe such an evaluation can only be task-based and would thus require the adaptation of an existing system (e.g., a syntactic parser) so it can benefit from our output.

12.3.3. Inferring Word Classes

In the first part of this dissertation, and in particular in chapter 5, we reached the conclusion that both a measure of *autonomy* and a procedure are needed to induce *wordclasses* in order to actually make the notion of *wordhood* computable. In the second part we addressed the question of the *autonomy* measure and we demonstrated how to use it to induce a sensible and ordered list of word candidates. The induction of the *wordclasses* of these candidates remains to be done.

We thought about using existing part-of-speech induction methods on our segmented corpus (Christodoulopoulos et al., 2010). We did a first experiment in which we followed Chrupała (2012) to obtain soft wordclasses and build an HMM from these classes and our segmentation. We could then run an Expectation Maximization (EM) algorithm to perform unsupervised POS tagging. Unfortunately, we ran into computation time issues due to the large size of our corpus, especially as we wished to take into account the n-best segmentations during the EM procedure.

We also consider using graph-based modelling of word candidates and their contexts (Kratochvíl's 1967 *substitution frames*) as a bigraph on which we can apply a chosen clustering method (Navarro, 2013) as this simple formulation would nicely match the initial idea.

Nevertheless, it is worth noting that, as expected from part 1 (especially sketched in chapter 5), the autonomy measure can not only be used to recognise word candidates but also to delimit the relevant contexts on which we can base a distributional

analysis. Considering the "mistakes" our system is likely to make (chapter 11), there are good reasons to think that a combination of two forms in which one is indissociable from the other will have a high autonomy as a compound even though the dissociable form is likely to be a word in itself (and thus will be present in our candidate list). Using this insight, we can find a way to focus on the indissociable component as a context for the dissociable one and this would fit our needs as presented in chapter 3.

12.4. Discussion

In this dissertation, we described the difficulties in defining a unit similar to the "word" in MSC linguistics and related this question to that of multi-word expressions regardless of the language. We argued that the issues which arise when defining the minimal units for linguistic description are not specific to Chinese. However particularities of MSC and of the Chinese script, as well as the availability of manually annotated corpora and a large body of literature on Chinese Word Segmentation designate MSC as a language of choice for our study.

It appeared to us that an important source of confusion for works in both Chinese Word Segmentation and Multi-word Expressions is the assumption of a correspondence between syntactic and semantic atomic units. The traditional orthography (the *sociological word*) also plays its part in the confusion. In any given language, the "correct" segmentation which follows the standard orthographic norm may not always be linguistically relevant.

We analysed the evolution of CWS as a task in NLP and pointed out that to satisfy the aforementioned assumption and ensure consistency of the segmentation, the authors of the various corpora had to define sets of heuristics which may conflict with the linguistic analysis and produce a somewhat arbitrary *orthography* (even though experiments tend to confirm its consistency).

This arbitrariness in the annotation schemes and the evaluation procedures tends to favour supervised machine learning. We think that the present supremacy of supervised learning techniques for CWS is grounded on linguistically irrelevant aspects of the datasets and that for the sake of linguistic description, there is more to expect from unsupervised learning and induction of structure from raw data.

We have not tried to propose a new definition of the segmentation unit or

new criteria to perform a segmentation which would clearly dissociate syntax and semantics from the start. This would have led us to graded phenomena difficult to cope with in an unquestionable way. We tried instead to describe the basic principle behind such a definition to open the way towards a purely data-driven, reproducible definition of wordhood. We concluded that the two main clues required for such a task are the combinatoric *autonomy* of a form and the possibility to assign it to at least one distributional class.

The second part of our work addressed the modelling of the first clue.

We started from previous works that tried to adapt to corpus studies an Harrissian hypothesis for morpheme segmentation. We refined its reformulation in order to best take into account the fact that our work is based solely on corpus data and a novel segmentation algorithm. By doing so we significantly improved on previous works based on the same hypothesis.

We also stressed the necessity to distinguish between "natural" and "unnatural" aspects of a language. Our *autonomy* measure concerns only the most "natural" parts of language. We call "unnatural" the factoid expressions that are created or understood by following simple rules that are often arbitrary and have to be learnt explicitly. For our study, they could be dealt with using simpler rule-based systems during a pre-processing step. This allowed us to enhance the quality of our segmentation of these phenomena and at the same time to discard irrelevant data that introduces noise into the estimation of the *autonomy* on the most "natural" parts of the language.

A last aspect of our work was to conduct a variety of experiments to demonstrate the genericity of our approach and its adequacy to our goal beyond the traditional evaluation of CWS. This led us to criticisms regarding the use of MDL for word segmentation. We also proposed a simple method to obtain a probabilistic model of autonomy.

In a future work, we shall reuse this probabilistic formulation of combinatoric *autonomy* with the detection of distributional classes based on the contexts of occurrences in order to obtain a more complete definition of wordhood.

It seems to us that the traditional insights used to define word classes on a distributional ground could be followed to model words and contexts as bigraphs on which community detections algorithm should provide a relevant classification. The

main challenge left is a proper definition of the relevant contexts, and the *autonomy* measure may provide an efficient filter.

In this work we avoided issues raised by inflectional morphology by focusing on MSC. In the future we may want to address these issues. Preliminary results on French show an over-segmentation of morphemes boundaries. We believe that analogical reasoning based on our probabilist lexicon may help us to detect inflectional paradigms and the associated indissociable affixes.

Appendix A

Typed Evaluation

Table A.1.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, with our basic system**

Accuracy	type	# occurrences
0.235	ancestor(epistemics,quantity)	183
0.361	ancestor(deontics,quantity)	266
0.383	ancestor(epistemics,evaluation)	324
0.433	ancestor(epistemics,epistemics)	127
0.443	ancestor(negation,evaluation)	553
0.446	ancestor(epistemics,time)	368
0.533	negation	2229
0.549	ancestor(deontics,epistemics)	133
0.572	ancestor(deontics,evaluation)	437
0.591	ancestor(deontics,time)	523
0.636	aspect	3544
0.642	quantity	2674
0.650	ancestor(quantifier,quantifier)	314
0.653	ancestor(negation,quantity)	121
0.680	degree	3199
0.692	evaluation	3282
0.723	ancestor(degree,evaluation)	173
0.727	predication	2488
0.750	concession	144
0.764	ancestor(negation,time)	212
0.778	ancestor(time,evaluation)	468

Table A.1.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, with our basic system (continued)**

Accuracy	type	# occurrences
0.778	quantifier	8540
0.781	ancestor(agent,evaluation)	178
0.795	complement	1510
0.796	time	6892
0.803	location	1699
0.811	ancestor(head,aspect)	127
0.818	property	33957
0.819	particle	1396
0.829	possessor	1690
0.837	ancestor(head,dummy1)	123
0.839	ancestor(deontics,agent)	360
0.840	ancestor(head,goal)	119
0.843	ancestor(goal,time)	172
0.845	ancestor(location,time)	341
0.849	ancestor(epistemics,agent)	252
0.851	contrast	161
0.854	deixis	727
0.858	purpose	141
0.858	ancestor(quantity,evaluation)	134
0.860	ancestor(location,agent)	243
0.862	ancestor(particle,complement)	123
0.863	experiencer	633
0.866	ancestor(time,time)	1055
0.869	head	16691
0.872	dummy	9005
0.874	theme	9191
0.877	ancestor(nominal,property)	457
0.879	deontics	2751
0.879	ancestor(degree,experiencer)	124
0.880	ancestor(quantity,experiencer)	125
0.881	goal	9545
0.883	ancestor(time,deontics)	180
0.885	ancestor(goal,agent)	131

Table A.1.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, with our basic system (continued)**

Accuracy	type	# occurrences
0.886	range	3742
0.887	ancestor(deontics,theme)	318
0.890	ancestor(epistemics,theme)	453
0.891	reason	339
0.892	addition	316
0.894	condition	321
0.898	ancestor(theme,addition)	128
0.899	ancestor(dummy1,property)	345
0.899	manner	4278
0.899	ancestor(manner,evaluation)	288
0.899	ancestor(time,experiencer)	159
0.899	hypothesis	189
0.901	agent	5273
0.903	ancestor(degree,time)	144
0.906	ancestor(dummy2,head)	117
0.912	ancestor(head,head)	611
0.912	epistemics	1517
0.913	ancestor(time,contrast)	161
0.919	nominal	1328
0.922	ancestor(theme,epistemics)	116
0.923	ancestor(manner,time)	570
0.923	duration	196
0.925	ancestor(location,theme)	239
0.930	ancestor(time,epistemics)	200
0.932	ancestor(quantity,agent)	441
0.933	ancestor(theme,evaluation)	282
0.933	ancestor(negation,theme)	283
0.933	ancestor(evaluation,experiencer)	195
0.933	ancestor(theme,time)	720
0.938	ancestor(nominal,agent)	194
0.938	ancestor(agent,time)	551
0.940	result	480
0.941	ancestor(theme,hypothesis)	119

Table A.1.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, with our basic system (continued)**

Accuracy	type	# occurrences
0.943	ancestor(time,manner)	122
0.945	ancestor(property,property)	7646
0.946	ancestor(manner,deontics)	315
0.946	ancestor(evaluation,evaluation)	186
0.948	ancestor(manner,epistemics)	174
0.950	ancestor(nominal,theme)	357
0.953	ancestor(particle,range)	339
0.953	ancestor(evaluation,time)	342
0.954	causer	130
0.956	dummy2	4000
0.956	ancestor(complement,goal)	1347
0.956	<-*-	15116
0.957	ancestor(theme,goal)	1481
0.958	ancestor(degree,theme)	733
0.958	ancestor(time,agent)	1594
0.958	ancestor(complement,theme)	406
0.959	ancestor(particle,theme)	123
0.960	ancestor(complement,location)	174
0.960	ancestor(particle,goal)	425
0.963	ancestor(complement,range)	570
0.965	ancestor(quantity,theme)	834
0.966	ancestor(property,quantifier)	1917
0.966	-*->	32239
0.967	ancestor(quantifier,property)	967
0.967	ancestor(time,theme)	1881
0.967	ancestor(time,location)	123
0.968	ancestor(manner,manner)	155
0.968	ancestor(theme,reason)	345
0.968	ancestor(evaluation,agent)	660
0.969	ancestor(theme,aspect)	703
0.969	apposition	1885
0.970	ancestor(property,predication)	873
0.971	ancestor(property,possessor)	382

Table A.1.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, with our basic system (continued)**

Accuracy	type	# occurrences
0.973	ancestor(complement,complement)	257
0.973	dummy1	4592
0.977	ancestor(quantifier,apposition)	216
0.980	ancestor(property,nominal)	152
0.982	ancestor(goal,aspect)	1261
0.982	ancestor(manner,theme)	453
0.983	ancestor(manner,agent)	632
0.983	ancestor(theme,theme)	873
0.983	ancestor(quantity,time)	241
0.984	ancestor(goal,goal)	1454
0.985	ancestor(predication,predication)	132
0.985	frequency	134
0.985	ancestor(theme,contrast)	405
0.986	ancestor(time,goal)	209
0.987	ancestor(quantifier,predication)	156
0.988	ancestor(dummy,dummy)	502
0.989	ancestor(range,range)	1226
0.991	ancestor(range,aspect)	323
0.991	ancestor(theme,location)	116
0.992	ancestor(time,topic)	121
0.993	ancestor(evaluation,goal)	141
0.993	ancestor(nominal,goal)	149
0.993	ancestor(theme,result)	152
0.995	ancestor(agent,agent)	188
0.995	ancestor(evaluation,theme)	1422
0.995	ancestor(property,apposition)	212
1.000	ancestor(agent,contrast)	241
1.000	ancestor(agent,result)	132
1.000	ancestor(theme,topic)	145

Table A.2.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, basic system+MDL**

Accuracy	type	# occurrences
0.213	ancestor(epistemics,quantity)	183
0.299	ancestor(epistemics,epistemics)	127
0.327	ancestor(deontics,quantity)	266
0.330	ancestor(epistemics,evaluation)	324
0.389	ancestor(epistemics,time)	368
0.421	ancestor(negation,evaluation)	553
0.466	ancestor(deontics,epistemics)	133
0.490	negation	2229
0.540	aspect	3544
0.545	ancestor(deontics,evaluation)	437
0.556	ancestor(deontics,time)	523
0.595	ancestor(negation,quantity)	121
0.600	degree	3199
0.611	quantity	2674
0.624	ancestor(quantifier,quantifier)	314
0.640	ancestor(agent,evaluation)	178
0.673	evaluation	3282
0.686	ancestor(goal,time)	172
0.687	quantifier	8540
0.705	ancestor(degree,evaluation)	173
0.712	ancestor(negation,time)	212
0.725	complement	1510
0.729	ancestor(time,evaluation)	468
0.739	deixis	727
0.756	ancestor(time,deontics)	180
0.756	location	1699
0.758	ancestor(degree,experiencer)	124
0.760	ancestor(quantity,experiencer)	125
0.763	ancestor(goal,agent)	131
0.764	time	6892
0.764	ancestor(head,aspect)	127
0.764	ancestor(evaluation,experiencer)	195

Table A.2.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, basic system+MDL (continued)**

Accuracy	type	# occurrences
0.770	ancestor(time,time)	1055
0.779	predication	2488
0.782	condition	321
0.785	ancestor(evaluation,evaluation)	186
0.789	ancestor(deontics,agent)	360
0.791	particle	1396
0.799	concession	144
0.802	ancestor(epistemics,agent)	252
0.803	dummy	9005
0.804	property	33957
0.814	contrast	161
0.816	ancestor(evaluation,time)	342
0.820	hypothesis	189
0.823	purpose	141
0.825	ancestor(time,epistemics)	200
0.826	ancestor(agent,result)	132
0.827	deontics	2751
0.829	possessor	1690
0.829	ancestor(head,dummy1)	123
0.829	ancestor(particle,complement)	123
0.836	ancestor(quantity,evaluation)	134
0.838	goal	9545
0.839	ancestor(location,time)	341
0.840	ancestor(head,goal)	119
0.846	ancestor(dummy2,head)	117
0.852	agent	5273
0.855	ancestor(agent,time)	551
0.858	ancestor(evaluation,agent)	660
0.858	experiencer	633
0.858	ancestor(deontics,theme)	318
0.859	manner	4278
0.859	ancestor(theme,addition)	128
0.862	ancestor(theme,epistemics)	116

Table A.2.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, basic system+MDL (continued)**

Accuracy	type	# occurrences
0.863	ancestor(time,contrast)	161
0.865	result	480
0.865	theme	9191
0.871	ancestor(quantity,agent)	441
0.871	range	3742
0.874	ancestor(time,experiencer)	159
0.885	ancestor(epistemics,theme)	453
0.885	ancestor(manner,evaluation)	288
0.886	ancestor(theme,time)	720
0.886	ancestor(nominal,property)	457
0.888	reason	339
0.889	head	16691
0.889	addition	316
0.891	epistemics	1517
0.894	ancestor(theme,evaluation)	282
0.897	ancestor(head,head)	611
0.897	ancestor(location,agent)	243
0.898	ancestor(manner,time)	570
0.902	ancestor(particle,theme)	123
0.902	ancestor(property,quantifier)	1917
0.903	duration	196
0.908	ancestor(theme,hypothesis)	119
0.910	nominal	1328
0.910	ancestor(time,manner)	122
0.910	ancestor(dummy1,property)	345
0.911	ancestor(manner,deontics)	315
0.912	ancestor(location,theme)	239
0.914	ancestor(complement,location)	174
0.915	ancestor(quantity,theme)	834
0.917	ancestor(degree,time)	144
0.917	ancestor(agent,contrast)	241
0.920	ancestor(particle,goal)	425
0.920	ancestor(particle,range)	339

Table A.2.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, basic system+MDL (continued)**

Accuracy	type	# occurrences
0.921	ancestor(quantity,time)	241
0.932	ancestor(time,agent)	1594
0.933	apposition	1885
0.933	ancestor(complement,range)	570
0.937	ancestor(manner,epistemics)	174
0.940	ancestor(negation,theme)	283
0.941	ancestor(theme,result)	152
0.942	ancestor(property,property)	7646
0.944	ancestor(complement,goal)	1347
0.945	ancestor(theme,reason)	345
0.948	ancestor(manner,manner)	155
0.950	dummy2	4000
0.951	ancestor(complement,theme)	406
0.951	ancestor(theme,goal)	1481
0.952	<-*-	15116
0.953	ancestor(theme,contrast)	405
0.954	-*->	32239
0.955	ancestor(degree,theme)	733
0.957	ancestor(complement,complement)	257
0.961	dummy1	4592
0.962	causer	130
0.963	ancestor(quantifier,property)	967
0.963	ancestor(theme,aspect)	703
0.963	ancestor(time,theme)	1881
0.967	ancestor(time,location)	123
0.969	ancestor(nominal,agent)	194
0.972	ancestor(evaluation,goal)	141
0.972	ancestor(property,apposition)	212
0.972	ancestor(quantifier,apposition)	216
0.972	ancestor(theme,topic)	145
0.973	ancestor(property,predication)	873
0.973	ancestor(evaluation,theme)	1422
0.974	ancestor(property,possessor)	382

Table A.2.: **Typed positions evaluation on boundaries, for types with at least 100 occurrences, basic system+MDL (continued)**

Accuracy	type	# occurrences
0.975	ancestor(goal,goal)	1454
0.975	ancestor(nominal,theme)	357
0.975	ancestor(goal,aspect)	1261
0.976	ancestor(theme,theme)	873
0.979	ancestor(agent,agent)	188
0.979	ancestor(manner,agent)	632
0.980	ancestor(property,nominal)	152
0.982	ancestor(manner,theme)	453
0.983	ancestor(range,range)	1226
0.985	ancestor(predication,predication)	132
0.985	frequency	134
0.986	ancestor(time,goal)	209
0.986	ancestor(dummy,dummy)	502
0.987	ancestor(quantifier,predication)	156
0.988	ancestor(range,aspect)	323
0.991	ancestor(theme,location)	116
0.992	ancestor(time,topic)	121
0.993	ancestor(nominal,goal)	149

Appendix B

French Data

Random sample of the French wikipedia from which we removed all the spaces before applying our segmentation system as proposed in chapter 8.

B.1. Using $N_\mu VBE$

Bretagne d' or et d'argent Les or f èvre sde ba sse Bretagne Si bi ri l(F in is tère), cha p elle re li qu aire de Saint- Ma ud ez, 14 47 La ri che sse del' or f èvre ri ebr et onne con s er vé ea inspir ésa présent ation end eux temps d ont le premier con c er ne lapartie occ id ent ale dela Bretagne ,di te ba sse Bretagne ou Bretagne "br et onnant e" . Parmiles trois c ent s oeuvre s déjà sé le c tion né es pour l' ex position ,la tr ent a ine publi é eici é vo que s ix siècle sde l' histoire de c et ar t pré ci eux . Entre le X I eet le X I I es ièccle ,d ont sub sistent de très r ar es in v ent aire sde tr és or s, les plus an ci ens ate li ers d' or f èvre s étaient , très pr ob able ment ,li és aux ate li ers monétaire sde R enne s, N ant es, Gu ing amp et V anne s, qui re le va i ent del' aut or ité du ca le. Les siège sd' évê ché s, T ré gui er et Qu imp er pour c it erles mi eux in form és, les ab ba y es im port ant es comme Saint- G il das -de- Rhuys ,on t eux aussi , con tribu é très tôt àla comm and eet àl' éla b or ation de pièce sd' or f èvre ri e. Aux X I V eet X Ve siècle s, les m en tion sde quelques no ms d' or f èvre s appar a issent ,a ins ique les premier spo in çon sde commun aut és, tel celuide M or la ix au milieu du X V e. Dans l' en semble ,né an moins ,le s con di tion sde l' exerc ic edu métier demeure nt mal connu es. Cette si tu ation se prolong edu r ant la première moitié du XVIe siècle et re fl ète alors l est â ton n ement s dela période de r attach ement dela Bretagne àla France . Après le tra ité d' Un ion de 15 32 et malgré la supp r es sion en 15 34 dela monnaie br et onne

,l' activité des or f èvre sse pour su it int ens ément enb asse Bretagne . En té mo ignent les diverses oeuvre s aux po in çon sà l'h er mi ne pas s ant e général is és ici comme enhaut e Bretagne au XVIe siècle . Tout efois , surla so ix ant a in ede pièce sde ba sse Bretagne con s er vé es de ce siècle ,un e par ti ene comp or te pas ce type de mar que ma is seulement le po in çon de m aît re , parfois difficile à identifi er ,en l' abs ence de statut s et d' ar chi ves relative s aux commun aut és. L'é re c tion des commun au té sd' or f èvre s en jur and es, permet t antun me ille ur c ont rôle du métier etdu titre du mé t al, est effect ive à N ant es età R enne s dès 15 79 ma is r es te difficile à dat er enb asse Bretagne . L' int ens e activité de M or la ix etl' utilis ation par cette v ille de po in çon sà lettre s- dat es dès 16 07 d onne àp ens er qu' elle était déjà les ième d'un e jur an de ,ce que confirm een 16 99 l' in spect ion des juge s dela M onna i ede R enne s. C'est en effet àl' extrême f in du X V I I es ième seulement que s' établi t défini t iv ement laré par ti tion des struct ur es ,à par tir des ressort sre spect ifs des hôtel sde s monnaie sde R enne s et de N ant es qui rég issent le travail des mé t aux pré ci eux . Ainsi , del' hôtel des monnaie sde R enne s dé p end ent enb asse Bretagne ,le s diocèse sde Saint- Po l- de - Léon et de T ré gui er, et enhaut e Bretagne ,c eux de Saint- Bri eu c, D ol, Saint- Ma lo et R enne s; del' hôtel de N ant es ,dé p end ent enb asse Bretagne ,le s diocèse sde Qu imp er et de V ann es et ,en h au te Bretagne ,ce lui de N ant es. En même temps ,le c ont rôle del' exerc ic edu métier se fait plus ri g our eux ,t ant sur le titre du mé t alp ré ci eux employé que sur le squ alité sre qui ses del' or f èvre qui , pour acc éd eràla maîtr is e, se voit dansl' oblig ation , désormais clairement spéc ifié e, de fourni r un che f d' oeuvre .

B.2. Using N_zVBE

Bretagne d' or et d'argent Les orfèvre sde b asse Bretagne Si bir il(Finistère), cha p elle reliquaire de Saint- Ma ud ez, 14 47 La richesse del' orfèvre ri ebretonne conserv ée a inspir ésa présent ation endeux temps dont lepremier concerne lapartie occidental e dela Bretagne ,di te b asse Bretagne ou Bretagne "br et onnant e". Parmiles trois c ent s oeuvre s déjà sélection né es pourl' exposition ,la tr ent a ine publié eici évoqu es ix siècle sde l'histoire deceta rt précieux . Entre le XIe etle XII esiècle ,dont subsist entdetrès rar es inventaire sde trésor s, les plusancien s atelier sd'orfèvre s étaient , très pro b ablement ,li és aux ateliers monétaire sde Rennes , N ant es, Gu ing amp et V anne s, qui re lev aient de l'autorité du ca le. Les siège sd' évêchés , Tréguier et Qu imp erpour cit erles mieux informé s,lesabb aye s important es comme Saint- Gildas

-de- Rhuys ,on t euxaussi , contribu é très tôt à lacommande et àl' é laboration de pièce sd'orfèvre r ie. Aux X I V eet X V esiècle s, les m ention sde quelques nom sd'orfèvre s ap paraissent , ainsique l espremiers poinçon sde communauté s, tel celuide Mor la ix aumilieudu X V e. Dansl' ensemble , néanmoins ,le s condition s del' exercice du métier demeurent mal connu es. Cette situation se prolonge durant lapremière moitié du XVIesiècle et reflète alors les tât onnement sde lapériode de r attach ementdela Bretagne à laFrance . Après le tra ité d'U nion de 15 3 2et malgré la sup pression en15 34 dela monnaie bre t onne , l'activité des orfèvre sse poursuit int ens ément enb asse Bretagne . En témoignent l esdiverses oeuvre s aux poinçon sà l'hermine passant e généralis és ici comme enhaut e Bretagne au XVIesiècle . Toutefois, surla soixant ain ede pièce sde b asse Bretagne conservées dece siècle , unepartie ne comporte pas ce typede marque mais seulement le poinçon de m aître , parfois difficile à identifier ,en l' absence de statut s et d'archives relative s aux communauté s. L'é r ection des communauté sd'orfèvre s en jurande s, permettant unmeilleur contrôle du métier etdu titre du mé t al, esteffect ive à N anteset à Rennes dès 15 79 mais res te difficile à dat er enb asse Bretagne . L' intense activité deMorlaix etl' utilis ationpar cetteville de poinçon sà lettre s- dat es dès 16 07 d onne àp ens er qu'elle était déjà le siège d'un e jurande ,ce que confirm een 16 99 l' inspection d esjuges dela Monnaie de R enne s. C'est eneffet àl' extrême f in duXVII esiècle seulement que s' établi t définitive ment laré partition des structure s, àpartirde s ressort s respect ifs d eshôtels d esmonnaies de Rennes et deNantes quiré g issent letravail d esmétaux précieux . Ainsi, del'hôtel d esmonnaies de Rennes dépendent enb asse Bretagne ,le s diocèse sde Saint- Pol -de- Léon etde T ré gui er,et enhaut e Bretagne ,c eux de Saint- Bri eu c, D ol, Saint-Malo et Rennes ; del'hôtel deNantes , dépendent enb asse Bretagne ,le s diocèse s deQuimper et deVanneset , enhaut e Bretagne , celuide N ant es. En même temps ,le contrôle del' exercice du métier se fait plus ri g our eux ,t ant sur letitre du métal précieux employé que sur l esqualités re qui ses del' orfèvre qui,pour accéder à lamaîtrise ,se voit dansl' obligation , désormais clairement spécifi ée, de fournirun che f d'oeuvre .

Bibliography

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In *Treebanks*, pages 165–187. Kluwer.

Arcodia, G. F. and Basciano, B. (2012). On the Productivity of the Chinese Suffixes -兒R,-化HUà AND -頭TOU. *Taiwan Journal of Linguistics*, 10:89–118.

Baayen, H. (1992). Quantitative aspects of morphological productivity. In *Yearbook of morphology 1991*, pages 109–149. Springer.

Baayen, R. H. (2001). *Word frequency distributions*, volume 18. MIT Press.

Bernstein-Ratner, N. (1987). The phonology of parent–child speech. *Children's language*, 6:159–174.

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, 3(8):294–301.

Bresnan, J. (1982). *The mental representation of grammatical relations*, volume 1. The MIT Press.

Bresnan, J. (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Roots: Linguistics in search of its evidential base*, pages 75–96.

Bresnan, J., Cueni, A., Nikitina, T., Baayen, R. H., et al. (2007). Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Chao, Y. R. C. and Yang, L. S. (1962). *Concise Dictionary of Spoken Chinese*. Harvard University Press.

Chen, K.-J., Huang, C.-R., Chang, L.-P., and Hsu, H.-L. (1996). Sinica corpus: Design methodology for balanced corpora. *Language*, 167:176.

Chen, P. (1999). *Modern Chinese: history and sociolinguistics*. Cambridge University Press.

Chen, R.-C. (2013). An Improved MDL-Based Compression Algorithm for Unsupervised Word Segmentation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 166–170. Association for Computational Linguistics.

Chen, S.-n. (1957). 現代漢語語法講話. 湖南人民出版社.

Chomsky, N. (1970). *Remarks on Nominalization*, page 314. Ginn.

Christodoulopoulos, C., Goldwater, S., and Steedman, M. (2010). Two Decades of Unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584. Association for Computational Linguistics.

Chrupała, G. (2012). Hierarchical clustering of word class distributions. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 100–104. Association for Computational Linguistics.

Church, K. (2013). How many multiword expressions do people know? *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2):4:12.

Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., and Billot, S. (2011). Intégrer des connaissances linguistiques dans un CRF: application à l'apprentissage d'un segmenteur-étiqueteur du français. In *TALN'2011 - Traitement Automatique des Langues Naturelles*, Montpellier, France. ATALA.

Courtois, B., Garrigues, M., Gross, G., Gross, M., Jung, R., Mathieu Colas, M., Silberztein, M., and Vivès, R. (1997). Dictionnaire électronique des noms composés DELAC: les composants NA et NN. *Rapport Technique du LADL*, 55.

Dai, J. X. (1992). *Chinese morphology and its interface with the syntax*. PhD thesis, Ohio state University.

DeFrancis, J. (1984). *The Chinese language: Fact and fantasy*. University of Hawaii Press.

Desalle, Y., Hsieh, S.-K., Gaume, B., and Cheung, H. (2010). Towards an automatic measurement of verbal lexicon acquisition: the case for a young children-vs-adults categorization in French and Mandarin. In *Proceedings of 24th Pacific Asia Conference on Language, Information and Computation: Workshop on Model and Measurement of Meaning (M3), Sendai, Japan*.

Dong, Z., Dong, Q., and Hao, C. (2010). Word Segmentation Needs Change–From a Linguist's View. In *Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 1–7.

Drillon, J. (1991). *Traité de la ponctuation française*. Gallimard.

Duan, H., Sui, Z., Tian, Y., and Li, W. (2012). The CIPS-SIGHAN CLP 2012 Chinese Word Segmentation on MicroBlog Corpora Bakeoff. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Duanmu, S. (1998). *Wordhood in Chinese*, volume 105, pages 135–196. Walter de Gruyter.

Emerson, T. (2005). The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 133.

Fan, J. (1958). Xing-ming zuhe jian 'de' zi de yufa zuoyong [The grammatical function of de in adjective-noun constructions]. *Zhongguo Yuwen*.

Feng, H., Chen, K., Deng, X., and Zheng, W. (2004). Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Fourtassi, A., Börschinger, B., Johnson, M., and Dupoux, E. (2013). Why is English so easy to segment? In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 1–10, Sofia, Bulgaria. Association for Computational Linguistics.

Frank, M. C., Sharon Goldwater, Thomas L. Griffiths, and Joshua B. Tenenbaum (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–125.

Gao, J., Li, M., Wu, A., and Huang, C.-N. (2005). Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics*, 31(4):531–574.

Goldwater, S. (2006). *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University.

Goldwater, S., Thomas L. Griffiths, and Mark Johnson (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Green, S., de Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 725–735. Association for Computational Linguistics.

Gross, G. (1996). *Les expressions figées en français: noms composés et autres locutions*. Editions Ophrys.

Gross, M. (1975). Méthodes en syntaxe. *Hermann, Paris, France*.

Gross, M. (1986). Lexicon-grammar: the representation of compound words. In *Proceedings of the 11th coference on Computational linguistics*, pages 1–6. Association for Computational Linguistics.

Harris, R. (2005). *Rethinking writing*. Continuum International Publishing Group.

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

Harris, Z. S. (1967). *Morpheme boundaries within words: Report on a computer test*. University of Pennsylvania.

Hewlett, D. and Cohen, P. (2011). Fully unsupervised word segmentation with BVE and MDL. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, volume 2, pages 540–545.

Hewlett, D. and Cohen, P. R. (2009). Bootstrap Voting Experts. In *IJCAI*, pages 1071–1076.

Hsu, C.-C. (2012). Lexical Gaps and Lexicalization: Implications for Word Segmentation Systems for Chinese NLP. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*. Faculty of Computer Science, Universitas Indonesia.

Huang, C. and Zhao, H. (2007). 中文分词十年回顾(Chinese word segmentation: A decade review). *Journal of Chinese Information Processing*, 21(3):8–20.

Huang, C.-R., Chen, F.-Y., Chen, K.-J., Gao, Z.-m., and Chen, K.-Y. (2000). Sinica Treebank: design criteria, annotation guidelines, and on-line interface. In *Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 12*, pages 29–37. Association for Computational Linguistics.

Huang, C.-R., Chen, K.-J., and Chang, L.-L. (1996). Segmentation standard for Chinese natural language processing. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1045–1048. Association for Computational Linguistics.

Huang, C.-T. J. (1984). Phrase structure, lexical integrity, and Chinese compounds. *Journal of the Chinese Language Teachers Association*, 19(2):53–78.

Iûnn, U.-G. (2007). New Manifestation of the Taiwanese vernacular literature – Introduction to Digital Archive for Written Taiwanese. In *National Museum of Taiwanese Literature*. NMTL.

Iûnn, U.-G., Tai, J.-h., Lau, K.-G., Chen, K., and Kao, C. (2009). Modeling Taiwanese POS tagging Using Statistical Methods and Mandarin Training Data. *International Journal of Computational Linguistics and Chinese Language Processing*, 14(3):237–256.

Jackendoff, R. (1997). *The architecture of the language faculty*. MIT Press.

Jespersen, O. (1924). *The Philosophy of Grammar*. Henry Holt.

Jin, G. and Chen, X. (2008). The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese pos tagging. In *Sixth SIGHAN Workshop on Chinese Language Processing*, page 69.

Jin, Z. (2007). *A Study On Unsupervised Segmentation Of Text Using Contextual Complexity*. PhD thesis, University of Tokyo.

Jin, Z. and Tanaka-Ishii, K. (2006). Unsupervised segmentation of Chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, page 428–435.

Johnson, M. and Demuth, K. (2010). Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 528–536, Beijing, China. Coling 2010 Organizing Committee.

Johnson, M. and Goldwater, S. (2009). Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325. Association for Computational Linguistics.

Johnson, M., Griffiths, T. L., and Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in neural information processing systems 19*.

Kahane, S. (2008). Les unités minimales de la syntaxe et de la sémantique: le cas du français. In *Congrès Mondial de Linguistique Française*. EDP Sciences.

Kaplan, R. M. and Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. *Formal Issues in Lexical-Functional Grammar*, pages 29–130.

Kaske, E. (2008). *The Politics of Language in Chinese Education, 1895–1919*, volume 82 of *Sinica Leidensia*. Barend J. ter Haar, Brill, Leiden, Boston.

Kempe, A. (1999). Experiments in unsupervised entropy-based corpus segmentation. In *Workshop of EACL in Computational Natural Language Learning*, page 7–13.

Klöter, H. (2005). *Written Taiwanese*. Harrassowitz.

Kratochvíl, P. (1967). Modern Standard Chinese. *Lingua*.

Levow, G.-A. (2006). The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, volume 117. Sydney: July.

Liu, F.-h. (1998). A clitic analysis of locative particles. *Journal of Chinese Linguistics*, 26(1):48–70.

Liu, F.-h. and Oakden, C. (2013). Disyllabic bound forms in Modern Mandarin Chinese: an analysis of *yihou* and *yihou*. *Journal of Chinese Linguistics*.

Lü, S. (1979). Hanyu yufa fenxi wenti [Issues in analysis of Chinese grammar].

Lu, Z. et al. (1964). Hanyu de goucifa [Chinese morphology]. *Revised edition. Beijing: Kexue Chubanshe*.

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

Magistry, P., Prévot, L., Cheung, H., Shiao, C.-y., Desalle, Y., and Gaume, B. (2009). Using Extra-Linguistic Material for Mandarin-French Verbal Constructions Comparison. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information, and Computation (PACLIC)*, pages 335–344.

Magistry, P. and Sagot, B. (2013). Can MDL Improve Unsupervised Chinese Word Segmentation? In *Sixth International Joint Conference on Natural Language Processing: 7th SIGHan Workshop*.

Martinet, A. (1960). *Eléments de linguistique générale*. Armand Colin, Paris.

Mathieu-Colas, M. (1988). Variations graphiques des mots composés dans le Petit Larousse et le Petit Robert. *Lingvisticae investigationes*, 12(2):235–280.

Mel'čuk, I. A. and Polguère, A. (1987). A formal lexicon in the Meaning-Text Theory: (or how to do lexica with words). *Computational linguistics*, 13(3-4):261–275.

Mel'čuk, I. (1988). *Dependency syntax: theory and practice*. State University Press of New York.

Mel'čuk, I. (1994). *Cours de morphologie générale: Significations morphologiques*, volume 2. les Presses de l'Université de Montréal.

Miller, P. H. (1992). *Clitics and constituents in phrase structure grammar*. Garland New York.

Mochihashi, D., Yamada, T., and Ueda, N. (2009). Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.

Navarro, E. (2013). *Métrologie des graphes de terrain, application à la construction de ressources lexicales et à la recherche d'information*. PhD thesis, University of Toulouse.

Nguyen, É. V. T. (2006). *Unité lexicale et morphologie en chinois mandarin: vers l'élaboration d'un dictionnaire explicatif et combinatoire du chinois*. PhD thesis, Université de Montréal (Canada).

Norman, J. (1988). *Chinese*. Cambridge University Press.

Packard, J. L. (2000). *The morphology of Chinese*. Cambridge University Press Cambridge.

Paris, M.-C. (1979). Some aspects of the syntax and semantics of the "lian...ye/dou" construction in Mandarin. *Cahiers de linguistique - Asie orientale*, 5(1):47–70.

Patin, G. (2013). *Extraction interactive et non supervisée de lexique en chinois contemporain appliquée à la constitution de ressources linguistiques dans un domaine spécialisé*. PhD thesis, Institut National des Langues et Civilisations Orientales.

Pearl, L., Sharon Goldwater, and Mark Steyvers (2010). How ideal are we? Incorporating human limitations into Bayesian models of word segmentation. In *Proceedings of the 34th annual Boston University Conference on Child Language Development*, pages 315–326, Somerville, MA. Cascadilla Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, 35(4):606–621.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Sagart, L. (2006). L'emploi des phonétiques dans l'écriture chinoise. In Bottéro et R. Djamouri, F., editor, *Ecriture chinoise/Données, usages et représentations*, Collection des Cahiers de Linguistique Asie Orientale, pages 35–53. Centre de Recherches Linguistiques sur l'Asie Orientale.

Sagot, B. and Boullier, P. (2005). From raw corpus to word lattices: robust pre-parsing processing with SXPipe. *Archives of Control Sciences*, 15(4):653–662.

Saussure, F. D., Bally, C., Sechehaye, A., and Riedlinger, A. (1916). *Cours de linguistique générale: publié par Charles Bally et Albert Sechehaye avec la collaboration de Albert Riedlinger*. Libraire Payot & Cie.

Sinclair, J. (1987). *Collins COBUILD, Collins Birgmingham University International Language Database: English language dictionary*. London [etc.]: Collins.

Sosa, A. V. and MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: collocations involving the word *of*. *Brain and Language*, 83(2): 227–236.

Sproat, R. and Emerson, T. (2003). The first international Chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 133–143. Association for Computational Linguistics.

Sproat, R., Gale, W., Shih, C., and Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. *Computational linguistics*, 22(3):377–404.

Sproat, R. and Shih, C. (1990). A Statistical Method for Finding Word Boundaries in Chinese Text. *Computer Processing of Chinese & Oriental Languages*, 4(4).

Suen, C. Y. (1986). *Computational studies of the most frequent Chinese words and sounds*. World Scientific Singapore.

Sun, J. and Lepage, Y. (2012). Can Word Segmentation be Considered Harmful for Statistical Machine Translation Tasks between Japanese and Chinese? In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC)*. Faculty of Computer Science, Universitas Indonesia.

Tanaka-Ishii, K. (2005). Entropy as an Indicator of Context Boundaries: An Experiment using a Web Search Engine. In *IJCNLP*, page 93–105.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klincksieck Paris.

Thuilier, J. (2012). *Contraintes préférentielles et ordre des mots en français*. PhD thesis, Université Paris-Diderot-Paris VII.

T'sou, B., Lin, H.-L., Liu, G., Chan, T., Hu, J., Chew, C.-h., and Tse, J. K. (1997). A synchronous Chinese language corpus from different speech communities: Construction and applications. *Computational Linguistics and Chinese Language Processing*, 2(1):91–104.

Walther, G. and Sagot, B. (2011). Problèmes d'intégration morphologique d'emprunts d'origine anglaise en français. In *30th International Conference on Lexis and Grammar*, Nicosia, Chypre.

Wang, H., Zhu, J., Tang, S., and Fan, X. (2011). A New Unsupervised Approach to Word Segmentation. *Computational Linguistics*, 37(3):421–454.

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

Xia, F. (2000). The segmentation guidelines for the Penn Chinese Treebank (3.0). Technical report, University of Pennsylvania.

Yu, H., Duan, S., Zhu, B., and Sun, X.-f. (2002). The Basic Processing of Contemporary Chinese Corpus at Peking University SPECIFICATION [J]. *Journal of Chinese Information Processing*, 5:007.

Zhang, Y. and Clark, S. (2007). Chinese segmentation with a word-based perceptron algorithm. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 840.

Zhang, Y. and Clark, S. (2010). A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, page 843–852.

Zhang, Y. and Clark, S. (2011). Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.

Zhao, H. and Kit, C. (2008). An empirical comparison of goodness measures for unsupervised Chinese word segmentation with a unified framework. In *The Third International Joint Conference on Natural Language Processing (IJCNLP2008), Hyderabad, India*.

Zhao, H. and Liu, Q. (2010). The CIPS-SIGHAN CLP 2010 Chinese word segmentation bakeoff. In *Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 199–209.

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

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley.

Zwicky, A. M. and Pullum, G. K. (1983). Cliticization vs. inflection: English n't. *Language*, pages 502–513.