Second Language Information Transfer in Automatic Verb Classification – A Preliminary Investigation

by

Vivian Tsang
曾婉莊

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Second Language Information Transfer in Automatic Verb Classification – A Preliminary Investigation

Vivian Tsang

Master of Science

Graduate Department of Computer Science

University of Toronto

2001

Lexical semantic classes incorporate both syntactic and semantic information about verbs. Lexical semantic classification of verbs provide a great deal of useful information about the possible usage of each verb. In our work, we explore the use of multilingual corpora in the automatic learning of verb classification. We extend the work of Merlo and Stevenson (2001a), in which statistics on simple syntactic features extracted from textual corpora were used to train an automatic classifier for three lexical semantic classes of English verbs. We hypothesize that some lexical semantic features which are difficult to detect superficially in English may manifest themselves syntactically in another language. In our two-way classification task, features from multiple languages achieve an accuracy as high as 81%, making a small bitext a useful alternative to using a large monolingual corpus for verb classification. In this thesis, experimental results are presented and future extensions are discussed.

Acknowledgements

First of all, I would like to thank my family. They have been extremely supportive of my decision to pursue a graduate degree. In particular, I would like to thank my mother, Annie Au-Yeung, and my grandmother, Po-po Cheung, who provided me with tremendous emotional and financial support throughout the last two years.

Next I must acknowledge Dr. Rena Helms-Park and her Ph.D. dissertation (1997). Although she and I only communicated briefly through emails, her thesis really sparked my interest in multilingualism and second language acquisition. I must also acknowledge Dr. Ron Smyth of the Linguistic Department here at the University of Toronto. Without his help, I would not have come across Helms-Park's work which led to this thesis.

I am extremely grateful to my supervisor, Dr. Suzanne Stevenson, whose work in automatic verb classification was the main inspiration for my attempt to bring the two fields, human and automatic language acquisition, together. In addition, I am indebted to her help in straightening out my thesis. Her super-human patience is what makes the completion of this thesis possible.

I thank the help from my second reader, Dr. Graeme Hirst, whose comments were critical but fair. Good critic makes the work stronger. I will certainly keep his comments in mind in my future writing.

On the coding side of things, I thank Eric Joanis for his help in the data extraction step.

Finally, I must also thank the following people who painstakingly went through various drafts of my thesis: my cousin Patrick Burkhalter, Afsaneh Fazli, Bowen Hui, Diana Zaiu Inkpen, Vikki Leung, and Vince Ma. Their comments are very much appreciated.

Contents

1 Introduction **1**

1.1 Statement of Purpose 1

1.2 Outline of the Study 4

2 Related Work **6**

2.1 Second Language Acquisition 6

2.2 Automatic Lexical Acquisition 9

2.3 Bringing the Two Areas Together 13

3 Chinese Features for Automatic Classification **15**

3.1 Linguistic and Methodological Assumptions 16

3.2 Materials 18

3.3 Feature Selection 19

3.3.1 Overview 19

3.3.2 Chinese POS tags for Verbs 22

3.3.3 (External) Periphrastic/Causative Particles 24

3.3.4 (External) Passive Particles 26

3.3.5 Morpheme Types 27

3.3.5.1 Compounding Pattern 27

3.3.5.2 Semantic Specificity 29

3.3.5.3 Average Morpheme Length 30

3.3.6 Summary of Features and Their Predicted Behaviour 31

4 Data Collection **33**

4.1 Materials and Method 33

4.1.1 Manual Corpus Analysis 34

4.1.2 Chinese Feature Extraction 35

4.1.2.1 Chinese POS tags 37

4.1.2.2 (External) Causative/Periphrastic Particles 37

4.1.2.3 (External) Passive Particles 38

4.1.2.4 Sublexical Information 38

4.1.3 English Feature Extraction 39

4.2 Data Analysis 40

4.2.1 Chinese Data 41

4.2.2 English Data 43

5 Experimental Results **46**

5.1 Experimental Methodologies 47

5.2 Results Using N-Fold Cross-Validation 48

5.2.1 HKLaws Data 50

5.2.2 WSJ Data 55

5.2.3 Summary of Cross-Validation Results 59

5.3 Results Using Leave-One-Out Methodology 60

5.4 Summary of Results 65

6 Discussion **68**

6.1 Chinese Lexical/Sublexical Features and Verb Classification 69

6.1.1 Individual Feature Performance 69

6.1.2 Contribution of Chinese Features to English Verb Classification . 71

6.2 The Use of Multilingual Corpora 74

7 Conclusions **76**

7.1 Using Multilingual Corpora for Automatic Learning in English 77

7.1.1 Contributions 77

7.1.2 Future Work: Corpus Size and Genre 78

7.2 Selection of "L1 Features" 78

7.2.1 Contributions 78

7.2.2 Future Work: Chinese Verb Classes and Interlingual Representations 79

7.3 "L1 Transfer" in Automatic Learning 81

7.3.1 Contributions 81

7.3.2 Future Work – Bi-directional Learning 81

A Chinese HKLaws Data **83**

A.1 Aligned Method 83

A.2 Unaligned Method 87

B English HKLaws Data **92**

C WSJ Data **94**

Bibliography **96**

List of Tables

2.1 Thematic role patterns by verb class 13

3.1 The Chinese features and their expected behaviour. 32

4.1 English unaccusative verbs and their corresponding Chinese verbs extracted from the HKLaws. 35

4.2 English object-drop verbs and their corresponding Chinese verbs extracted from the HKLaws. 36

4.3 Manually and automatically calculated Chinese feature frequencies of a random sample of verbs, aligned method 41

4.4 Manually and automatically calculated Chinese feature frequencies of a random sample of verbs, unaligned method 42

4.5 Manually and automatically counted English feature frequencies for a random sample of verbs extracted from the HKLaws 43

4.6 Manually and automatically counted English feature frequencies for a random sample of verbs extracted from the WSJ 45

5.1 Percent accuracy and standard error of the verb classification task using each English feature individually, with 8-fold cross-validation training method repeated 50 times. 50

5.2 Percent accuracy and standard error of the verb classification task using each Chinese feature individually, with 8-fold cross-validation training method repeated 50 times. 51

5.3 Percent accuracy and standard error of the verb classification task by removing each individual English feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times. 51

5.4 Percent accuracy and standard error of the verb classification task by removing each individual Chinese feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times. 52

5.5 Percent accuracy and standard error of the verb classification task using all English-only features, all Chinese-only features, and the full bilingual feature set, with 8-fold cross-validation training method repeated 50 times. 53

5.6 Percent accuracy and standard error of the verb classification task by augmenting the full English feature set with each individual Chinese feature, with 8-fold cross-validation training method repeated 50 times. 53

5.7 Percent accuracy and standard error of the verb classification task using each English feature individually, with 8-fold cross-validation training method repeated 50 times. English WSJ data used. 55

5.8 Percent accuracy and standard error of the verb classification task by removing each individual English feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data. 55

5.9 Percent accuracy and standard error of the verb classification task by removing each individual Chinese feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data. 56

5.10 Percent accuracy and standard error of the verb classification task by augmenting all the English features with each individual Chinese feature, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data. 57

5.11 Percent accuracy and standard error of the verb classification task using all English-only features, all Chinese-only features, and all features, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data. 58

5.12 Percent accuracy and standard error of the verb classification task by augmenting all the English features with each individual Chinese feature, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data. 58

5.13 Recall, precision, balanced F-score, and percent accuracy of the verb classification task using all Chinese features, with leave-one-out training method. 62

5.14 Recall, precision, balanced F-score, and percent accuracy of the verb classification task using all English features and a combination of Chinese features, with leave-one-out training method. (1 = CKIP Tags; 2 = Passive Particles; 3 = Periphrastic Particles) 63

5.15 Recall, precision, balanced F-score, and percent accuracy of the verb classification task using all English features - Animacy + a combination of Chinese features, with leave-one-out training method. (1 = CKIP Tags; 2 = Passive Particles; 3 = Periphrastic Particles) 64

5.16 Summary of the best feature combinations using N-fold cross-validation training methodology 65

5.17 Summary of the best feature combinations using leave-one-out training methodology (1 = CKIP Tags; 2 = Passive Particles; 3 = Periphrastic Particles) 66

6.1 The Chinese features and their expected behaviour. 69

Chapter 1

Introduction

1.1 Statement of Purpose

The aim of this study is to examine the use of multilingual resources in the automatic learning of verb classification. In recent years, researchers have focused on collecting statistical distributions of argument structures and diathesis alternations (different ways arguments of a verb are expressed) (e.g., Lapata and Brew, 1999; McCarthy, 2000; Merlo and Stevenson, 2001a). These works have confirmed Levin's (1993) central idea that some semantic properties of a verb are expressed syntactically. Further, they have shown that corpus statistics on a set of carefully selected classes of verbs can capture information about each class of verbs.

We would like to extend this type of corpus-based technique by exploring the use of bilingual corpora (bitext), parallel and otherwise. Bitext-based techniques have been used for various natural language processing (NLP) tasks (e.g., Gale and Church, 1991; Fung, 1998; Melamed and Marcus, 1998; Pao, 2000). However, many of them focus on corpus alignment or mining translational equivalents; few consider the automatic learning of a property of one language using data from another language. We refer to this idea generally as "information transfer" even in instances where the data from the

second language may prove to be not useful. That is, any information that is not already captured by the data in the first language is considered as a "transfer" from one language to another.

Consider the following example of a passive voice sentence construction in English:

1. This store is closed by the owner.

The above English sentence is detected as passive by observing the *be*-auxiliary verb, *is*, preceding the main verb, *closed*, in the past participle form. Although with a simple POS tagger (part-of-speech tagger), we are able to detect that the above English sentence is in the passive voice, not all English passive voice sentences are easy to detect, even by humans. Consider this "garden path" example:

2. The doctor called to the scene examined the injured man.

This sentence contains a reduced relative clause *called to the scene*. One interpretation would be "The doctor *who was called to the scene* examined the injured man", in which the relative clause is a passive voice clause. However, *call* is a regular verb – its past tense form is exactly the same as its past participle form – which causes a possible local ambiguity. The preferred parse at the word *called*, i.e., the intransitive sentence "The doctor called", is not necessarily the correct parse for the whole sentence (hence, the name "garden path"[1]). Following the same argument, current automatic annotation methods might not process these types of sentences correctly.

Now let's consider the Chinese[2] equivalent of sentence 1:

[1]However, not all "garden path" sentences are difficult to comprehend. In sentence 2, the local ambiguity could easily be resolved at the end of the relative clause. However, it is not the case for some of the classic garden path examples, such as "The horse *raced past the barn* fell." It is suggested that the ease/difficulty of parsing such garden path sentences depends on the lexical semantic membership of the verb in the reduced relative clause. Please refer to (Stevenson and Merlo, 1997) for a detailed discussion.

[2]In this thesis, we use "Chinese" to refer to Mandarin Chinese. Other dialects of Chinese are not considered here.

3. 這個 商店 被 東主 關閉。
   This store bei (passive particle) owner closed
   This store is closed by the owner.

The Chinese sentence is detected as passive by observing the overt preverbal passive particle *bei* (被). Similarly, we see the use of the same passive particle for the Chinese translation of sentence 2:

4. 被 召喚 到達現場 的
   bei (passive particle) called to the scene de (adjectival particle)
   醫生 檢查 受傷的男人。
   doctor examine the injured man
   The doctor (who was) called to the scene examined the injured man.

Observe that the relative clause in sentence 2 is moved to the beginning of the translated sentence. Since there is no relative clause construction in Chinese, the grammatical function of the clause is changed into an adjective phrase as indicated by the adjectival particle *de* (的).

These examples show that it is not always easy to detect English syntactic constructions (such as the passive construction in sentence 2), but the equivalent features may be obvious in the translation in another language such as Chinese. If we have access to a parallel corpus in, say, English and Chinese, the concern of parsing sentence 2 correctly would be a non-issue. This leads us to believe that we can benefit from exploiting non-English data for some corpus-based, automatic learning tasks in English.

The idea of "information transfer" is not new, especially in areas such as the study of Second Language Acquisition (SLA). As the name suggests, SLA research studies how humans acquire a new language (L2). One complicating issue is that **prior knowledge** of the first language (L1) can affect the acquisition of L2. We will not address the issue of how the established L1 lexicon (classifications and lexical rules) interacts with

the emerging L2 lexicon, nor will we address the nature of the organization of lexical knowledge. What is of interest to us is how the idea of L1 and L2 lexical interaction can carry over to the machine-learning setting. In particular, our goal is to use data in one language to aid the learning of verb classification in another language. We will describe our notion of information transfer in more detail in subsequent chapters.

1.2 Outline of the Study

In this study, English and Chinese are chosen to test the idea of information transfer in automatic learning. We chose English as one of the two languages because there are numerous existing studies on lexical semantic acquisition in English. We chose Chinese simply because the author is a native speaker of the language. We used two corpora, one monolingual (English) and one parallel (English-Chinese). The purpose of using a monolingual English corpus is twofold: first, to attempt to duplicate existing research in English; and second, to explore whether multilingual data obtained from separate non-alignable monolingual corpora is useful or not. Our aim is to compare the multilingual results with the monolingual results. More specifically, we have the following hypotheses:

- Based on existing research in English automatic learning tasks, monolingual English data (statistics of various syntactic features, extracted from a textual corpus) is useful in classifying English verbs.

- There are English features which are useful in dividing verbs into semantic classes, but which are difficult to detect syntactically (see our examples above). Hence, they are difficult to detect automatically in a corpus.

- There are Chinese features, whether they are linguistically related to these English features or not, that can be easily detected syntactically from a corpus.

- These Chinese features contribute to learning the classifications of English verbs.

- Using Chinese data in conjunction to English data aids the same classification task.

To test our hypotheses, we will collect data from each monolingual corpus/subcorpus: English data from the monolingual corpus, English data from the English subcorpus of the bitext, and Chinese data from the Chinese subcorpus of the bitext. Using this data, we will conduct experiments in the following contexts:

- Monolingual experiments: For each of the three sets of data, a verb classification task will be performed independently.
- Multilingual experiments: For each pair of English and Chinese data, we will perform the same verb classification task.

In the subsequent chapters, we will describe our experimental settings, report on our results, and demonstrate that our hypotheses are confirmed. In Chapter 2, we provide a survey of related work in SLA research and in automatic lexical acquisition. In Chapter 3, we present our linguistic assumptions and research methodologies. A contrastive analysis of some English and Chinese sentence constructions and a description of our Chinese features are also provided in the same chapter. In Chapter 4, we show how statistics are collected on our selected features. Results are reported in Chapter 5. Our results confirm that the classification performance using multilingual resources is comparable to the performance using monolingual data, which we will discuss in Chapter 6. In the last chapter, we conclude by discussing the contributions and limitations of our current research methodologies, and suggest possible future extensions.

Chapter 2

Related Work

2.1 Second Language Acquisition

Although this work is primarily a study in the use of multilingual information in automatic acquisition of lexical knowledge, the inspiration comes from SLA research on the role of L1 transfer. As mentioned in the introduction, one important branch of SLA research studies the organization of the L2 lexicon and the influence of prior L1 knowledge. If an L1 property is a possible source of error in the learner's L2, the transfer is considered as "negative". For example, in German, to express a change of state or change of location in the perfect tense, *be*-auxiliary verbs are used instead of *have*-auxiliary verbs:

1. (a) Klaus und ich *sind* nach Hause gegangen.
   Klaus and I are to home gone.
   (b) Klaus and I *have* gone home.
   (c) * Klaus and I *are* gone home.

It is likely that the *be*-auxiliary usage in German is a source of error in using English among native German speakers. However, if a learner is helped by a certain feature in L1 that exists in L2 as well, the transfer is "positive". For example, German speakers should have little problem acquiring the Dutch equivalent of the above sentence:

2. Klaus en ik *zijn* naar huis gegaan.
   Klaus and I are to home gone.

Many studies have documented the effects of L1 transfer (e.g., Wang and Lee, 1999; Yuan, 1999). The recent trend is to focus on the cross-linguistic differences in verb morphosyntax and semantics. This is largely influenced by Pinker (1989), Levin (1993), and Levin and Rappaport Hovav (1995) who suggest that there is a predictable link between a verb's meaning and its morphology and/or syntax. For example, both *tell* and *whisper* take one direct object and one indirect object.

3. (a) Mary tells John a secret.
   (b) Mary tells a secret to John.

4. (a) * Mary whispers John a secret.
   (b) Mary whispers a secret to John.

Although both verbs are communication verbs, only sentence 4a is problematic. That is, the verb *whisper* does not allow the double-object dative subcategorization frame. Pinker (1989) suggested that the two dative alternations are slightly different in meaning. The double-object dative construction "X Verb Z Y" means "X causes Z to have Y" (e.g., *Mary* causes *John* to have *a secret*.) The prepositional dative construction "X Verb Y to Z" in fact means "X causes Y to go to Z" (e.g., *Mary* causes *a secret* to go to *John*.) Sentence 3a is allowed because John knows a secret as a result of Mary's telling it. However, the same inference cannot be made for Mary's whispering action. That is, her whispering action does not always result in John knowing a secret (e.g., there is too much background noise or Mary does not whisper loud enough.) Pinker (1989) argued that the *whisper* class of verbs is different from the *tell* class of verbs because the *tell* class lacks the "manner" component (the "whispering" manner) of the meaning. The difference underlies the property that *tell* class verbs are grammatical in both dative constructions while the *whisper* class only allows the prepositional dative construction.

Interestingly enough, in Chinese, the same distinction exists between the *whisper* class and the *tell* class:

5. (a) 瑪利 告訴 約翰 一個秘密。
Mary tells John a secret

(b) 瑪利 告訴 一個秘密 給 約翰 聽。
Mary tells a secret to John listen

6. (a) * 瑪利 小聲說 約翰 一個秘密。
Mary whispers John a secret

(b) 瑪利 小聲說 一個秘密 給 約翰 聽。
Mary whispers a secret to John listen

Inagaki (1997) observed this same verb class distinction in both Chinese and English and hypothesized that Chinese learners of English should be able to distinguish the double-object dative constructions containing *tell* class verbs from those containing *whisper* class verbs; his results confirmed this prediction.[1] In fact, some Chinese subjects reported substituting English words with Chinese in a sentence rating task. The L1-based "substitution" strategy seems to help the Chinese subjects to distinguish the two classes of verbs. Inagaki's approach here is that given a fixed L2 (in this case, English), subjects are grouped by their L1s. In a grammatical judgement task, the groups' performance in one or more classes of English verbs are compared. The goal is to observe and compare any L1 transfer effects in different L1-L2 pairs.

During the 1990s, the effects of L1 transfer have been studied in many other L1-L2 pairs. The prevalent view is that L1 transfer effects are only prominent in the early stages

[1]Inagaki made a similar prediction about Japanese speakers' ability to distinguish *throw* class verbs from *push* class verbs in double-object constructions. However, his hypothesis does not apply to Japanese speakers. He postulated that the subject-verb-object (SVO) word order shared by Chinese and English facilitates L1 transfer more efficiently than Japanese, which is an SOV language.

of acquisition.[2] For example, Montrul (2000) found little L1 influence in L2 learners of Turkish, Spanish, and English with intermediate proficiency; Wang and Lee (1999) found that the wide range of Chinese adjectivizable verbs causes low proficiency Chinese learners to overgeneralize adjectivizable English verbs. Contrary to this view, we discovered one study focusing on the L1 transfer effects at later stages of acquisition: Helms-Park (1997) looked at the acquisition of several English verb classes that undergo causativization by Vietnamese and Hindi-Urdu speakers at various proficiency levels. The author found that, at a high proficiency level, Vietnamese speakers were able to unlearn the overuse of direct causatives sooner than Hindi-Urdu speakers, especially in classes such as directional motion, animal sounds/internal mechanisms, and emission. She concluded that L1 transfer is responsible for the unlearning since Vietnamese has a more limited range of direct causatives.

Obviously, L1 transfer is not the only phenomenon in SLA. However, our focus is to explore the use of multilingual data rather than the nature of the organization of the L1 and L2 lexicon. Descriptions of other aspects of SLA are omitted here. For a comprehensive review of the verb class approach in SLA, see (Juffs, 2000).

2.2 Automatic Lexical Acquisition

In this section, we choose to focus on related work in automatic acquisition of verb classes and alternations using large textual corpora. Specifically, we want to look at some of the same issues in human language learning, such as the importance of subcategorization frames in acquiring the meaning of verbs. Corpus-based methods have been widely used for various automatic lexical acquisition tasks (e.g., Oishi and Matsumoto, 1997; Lapata and Brew, 1999). However, these tasks are largely monolingual. One exception is the use

[2]People who are in their early stages of L2 acquisition tend to be at a lower proficiency level in L2 (e.g., as determined by standard language proficiency tests). Similarly, high proficiency learners of L2 tend to have been exposed to L2 for a longer period of time. We use the terms *early/late stages of acquisition* and *low/high proficiency* interchangeably.

of a parallel corpus for word sense disambiguation in the work of (Resnik and Yarowsky, 1997, 1999; Ide, 1999, 2000). These authors believe that a parallel English–non-English corpus should provide a source for lexicalizing some fine-grained English senses. Another possible use of multilingual resources is mentioned as future work by Siegel and McKeown (2000), who suggest parallel bilingual corpora may be useful in learning the aspectual classification (i.e., state or event) of English verbs. Finally, we want to mention the work by Aone and McKee (1996), in which the authors devise a predicate-argument extraction technique that works cross-linguistically (English, Spanish, and Japanese). Although they claim they were able to generalize the extraction process of syntactic and semantic features (such as subject animacy and transitivity) across different languages, the data collected is language-specific. That is, an extracted feature in one language is only useful in the same language. In essence, Aone and McKee's (1996) work, though done in multiple languages, is also monolingual.

There are similar verb classification tasks done in languages other than English. We discovered one corpus-based task classifying Japanese verbs. Using a Japanese corpus provided by the Japan Electronic Dictionary Research Institute Ltd., Oishi and Matsumoto (1997) classified Japanese verbs by detecting surface case markers (nominative, accusative, etc.) and co-occurring adverbials. The detection of these surface indicators gives rise to two orthogonal dimensions: the thematic dimension (surface argument structures) and the aspectual dimension (state, process, or transition). The result of their experiment was a coarse-grained semantic classification of Japanese verbs using Lexical Conceptual Structure (LCS). Although it is unlikely that Chinese verbs could be classified in the exact same way, we use an approach similar to that of Oishi and Matsumoto (1997) by devising a method that exploits Chinese surface particles, which we will return to in Sections 3.3.3 and 3.3.4.

We will now take a slight change in direction by focusing on the learning of subcategorization frames and diathesis alternations. Some SLA research considers subcat-

egorizations and diathesis alternations important in acquiring the syntactic frames and the meaning of verbs (e.g., Helms-Park, 1997; Inagaki, 1997). They can also be useful in automatic lexical acquisition. Korhonen (1997) used statistical methods to acquire subcategorization frames from the Susanne corpus. McCarthy and Korhonen (1998) and McCarthy (2000) used the acquired subcategorization information to learn diathesis alternations with the help of an external resource, WordNet. It is hypothesized that by using selectional preferences as a similarity measure, verbs which participate in the same alternation have similar arguments. For example, in the conative alternation (e.g., *The boy pulled at the rope/The boy pulled the rope* (McCarthy, 2000)), the selectional preferences for the object in a transitive use and the object in a prepositional use will show a high degree of similarity because the objects in the two constructions tend to be the same underlying argument. The selectional preferences for the transitive object and the prepositional object of verbs which do not participate in the conative alternation will be less similar because they are not the same underlying argument. In comparison to the random baseline accuracy of 50%, their method yields a best (mean) accuracy of 73% for the causative alternation and 67% for the conative alternation, where the improvements are statistically significant.

Subcategorization information is also helpful in Lapata and Brew's (1999) work in determining verb class membership based on Levin's (1993) classification. In Levin's (1993) verb index, close to 800 verbs belong to two or more semantic classes. Lapata and Brew (1999) found that the frequency distributions of subcategorization frames within and across classes are useful in disambiguating these polysemous verbs. Their model was tested on two different categories of verbs: verbs that can be disambiguated by their syntactic frame only, and verbs inhabiting one syntactic frame with multiple class memberships. For the former class of verbs, their method achieves an overall precision of 91.8% (compared to a baseline performance of 61.8%). For the ambiguous class of verbs, their method achieves an overall precision of 83.9% (baseline performance of 61.3%).

Merlo and Stevenson (2001a) used surface syntactic indicators to classify verbs in each of the three optionally transitive classes: unergative (manner-of-motion), unaccusative (change-of-state), and object-drop (a variety of optionally transitive classes from (Levin, 1993)) verbs.

Unergative: 7. (a) The jockey raced the horse past the barn.
(b) The horse raced past the barn.

Unaccusative: 8. (a) Mary melted the chocolate bar.
(b) The chocolate bar melted.

Object-Drop: 9. (a) The boy played soccer.
(b) The boy played.

These classes share the same subcategorizations, that is, they can be either transitive or intransitive, as seen in the above three examples. Thus, collecting subcategorization information is not sufficient to distinguish one verb class from the other. However, there are some differences in the thematic role assignments across the classes, as shown in Table 2.1. Observe that the three classes can be uniquely identified, despite sharing the same two subcategorizations. To identify these patterns, Merlo and Stevenson (2001a) derived the five syntactic features: transitivity, passive voice, past participle form, causativity, and animacy of the subject. The features are represented as a vector of normalized frequencies of usage of each verb in a corpus. It is suggested that these features are useful in approximating the statistical distribution of the thematic role assignments and thereby distinguishing the three classes. Consider the feature, animacy of the subject. Observe that only unaccusative verbs take a theme subject in the intransitive. Both unergative verbs and object-drop verbs assign an agentive role to their subject in both the transitive and the intransitive. Assuming (i) that unaccusatives occur frequently in the intransitive construction and (ii) agents are usually animate entities, there is a difference in the frequencies of having an animate subject between the verb classes – it is hypothesized

Class	Transitive		Intransitive
	Subject	Object	Subject
Unergative	Causative agent (e.g., the jockey)	Agent (e.g., the horse)	Agent (e.g., the horse)
Unaccusative	Causative agent (e.g., the cook)	Theme (e.g., the chocolate bar)	Theme (e.g., the chocolate bar)
Object-Drop	Agent (e.g., the boy)	Theme (e.g., soccer)	Agent (e.g., the boy)

Table 2.1: Thematic role patterns by verb class

that unaccusative verbs occur less often with an animate subject compared to the other two classes. This is one example of a feature for which the verb classes exhibit different linguistic behaviours. Such linguistic differences are manifested syntactically, and furthermore, there are different frequencies of usage of the syntactic behaviour.

Although we will discuss briefly the linguistic motivations behind Merlo and Stevenson's selection of these features in Chapter 3, specifically those that are linguistically related to our Chinese features, we will omit most of their linguistic analysis. We point the interested reader to (Merlo and Stevenson, 2001a) for further details.

Merlo and Stevenson (2001a) selected 59 verbs for their classification task, 19 verbs in one class, 20 verbs in each of the remaining two classes. The random baseline performance is 33.9% (one out of three classes or 20 out of 59 verbs). Using a combination of the above features, the best accuracy was as high as 71.2% on all verbs, and as high as 85% on the individual verb classes. The reported improvements are statistically significant.

2.3 Bringing the Two Areas Together

Much SLA research draws on the cross-linguistic differences in the acquisition of semantic verb classes. Verb classification criteria such as subcategorization frames or aspectual dimension are not the only criteria in human acquisition of verbs. We find that the learn-

ing approaches adopted by Merlo and Stevenson (2001a) target the level of granularity closest to our interest. For example, the authors extend the different learning approaches in the previous section. They improve on Oishi and Matsumoto's (1997) method of learning argument structure properties without resorting to external resources such as a dictionary. Further, their approach shows a finer-grained learning of verbs than those of McCarthy and Korhonen (1998) and McCarthy (2000) by making a distinction between the induced action alternation and causative/inchoative alternation (they share the same two syntactic frames), though their method of extracting the causativity feature is similar to McCarthy's (2000) way of calculating "lemma overlap".

We believe the work of Merlo and Stevenson (2001a) is closely related to the research on human language acquisition. For example, real-life data has shown that although children or L2 learners initially only acquire argument structures and alternations given positive evidence; during subsequent stages of acquisition, they generalize the alternations of other verbs (e.g., Pinker, 1989; Helms-Park, 1997). This is not unlike the approach taken by Merlo and Stevenson (2001a): they attempt to train an automatic classifier by collecting argument structure data (positive evidence) and apply the classifier to previously unseen verbs (generalization). We are aware that developing machine-learning techniques is not the same as understanding the human acquisition process, but we believe that we can benefit from the ideas in SLA research. For the purposes of our work, we choose to extend the work of Merlo and Stevenson (2001a) by using English as well as non-English (in our case, Chinese) data.

Chapter 3

Chinese Features for Automatic Classification

In Chapter 2, we have described the use of linguistic features in automatic verb classification. These features capture syntactic and semantic differences between verb classes. The differences are reflected in the amount of usage across the classes. Using a machine-learning algorithm, the frequency patterns are useful to identify the verb classes. In this chapter, we describe a similar selection process for Chinese features for automatic verb classification. Given an English-Chinese bilingual corpus, we suggest augmenting Merlo and Stevenson's (2001a) set of English verb features with a set of Chinese verb features, with the Chinese features carefully selected according to the different linguistic behaviour of verbs. The hypothesis that a non-English verb feature set will be useful stems from various studies in verb class acquisition among English-as-a-Second-Language (ESL) learners as described in the previous chapter (e.g., Helms-Park, 1997; Inagaki, 1997; Juffs, 2000). If one's native language influences human learning of (English) verbs, what lexical features are used? Do they affect the automatic acquisition of verb classes? If so, are they a hindrance or do they improve the automatic acquisition process?

3.1 Linguistic and Methodological Assumptions

Since we are extending the work by Merlo and Stevenson (2001a), we will focus on the same three optionally transitive verb classes: unergative (manner-of-motion) verbs, unaccusative (change-of-state) verbs, and object-drop verbs. We restrict the last to creation and transformation verbs from (Levin, 1993). To find a set of Chinese features, we first examine some samples of English-Chinese bitext documents, paying special attention to these three verb classes of interest. One good source of bitext is the Legislative Council website for the Hong Kong Special Administrative Region of the People's Republic of China (`http://www.legco.gov.hk`). Upon a quick inspection of a few council meeting documents in the year 2000-2001, we have the following observations. In a bitext, very rarely is an English verb mapped into one single Chinese verb with exactly the same meaning. Often it is translated into one or more of the following:

- a verb preceded by an external particle indicating causativity or passive voice (the verb alone does not indicate causativity or passiveness)

  e.g., The external particle *jiang* (將) is needed to indicate causativity.

  火 將 冰塊 溶解。

  fire make (particle) ice cubes melt

  Fire melts the ice cubes.

- a serial verb compound (multiple morphemes and/or verbs concatenated together to approximate the meaning)

  e.g., Neither 進 (means "go forward") nor 行 (means "in progress") means *perform*. However the compound 進行 can be interpreted as *perform* depending on the context.

- a verb that has a finer-grained meaning encoded in one or more morphemes

  e.g., 滴乾 is a translation for *drain*. However, 滴乾 carries a connotation that the result of the draining is dry which is not always the case for *drain*.

Some of these Chinese translations may encode features that are not readily observed as a surface feature in English. If there exists such a set of features, we want to find out if they shed light on the finer-grained distinctions between the three classes of optionally transitive English verbs. Before we address the issue of selecting Chinese features, we note the following assumptions we have made:

Chinese word segmentation We rely on a Chinese POS tagger to determine the word boundary because the notion of a word is fuzzy in Chinese. A Chinese character/morpheme is considered an atomic unit, which may itself be a "word unit". However, each morpheme can be concatenated with one or more morphemes to form a compound which can be considered another "word unit" with the same or a different meaning. A Chinese sentence is just a string of these morphemes and compounds concatenated together without any whitespace in between. The POS tagger we use does not provide perfect word segmentation. However, there is no agreement on the word segmentation issue even among native speakers. We consider the accuracy of the automatic segmentation sufficient for our task.

Chinese word order English and Chinese have similar subject-verb-object (SVO) word order. Clearly, there are some exceptions. Consider the example using a directed motion verb, *sink*.

在 這個 海域 沉了 三艘船。
in this sea area sank three ships
Three ships sank in this sea area. (Yuan, 1999)

Observe that in English, the argument, *three ships*, must be moved to preverbal subject position, but in Chinese, it can remain in the object position as long as the argument is an indefinite NP. That said, in Chinese, it is equally legal to move the argument to the subject position. Our example aside, we made the simplifying assumption that at least

in a parallel corpus, Chinese translations of English sentences retain the same SVO word order.

3.2 Materials

Our source of data is the *Hong Kong Laws Parallel Text* (HKLaws). It is a sentence-aligned bilingual corpus with 313,659 sentences per language, approximately 6.5 millions words in the English subcorpus, and 9 million characters in the Chinese portion. This corpus was obtained from the bilingual website of the Department of Justice of the Hong Kong Special Administrative Region of the People's Republic of China during the month of January 1999. Although traditionally all originals of government laws and press releases were in English, as early as 1989, both English and Chinese versions were considered "equally authentic" by the Hong Kong government (`http://www.justice.gov.hk`). Therefore, we can no longer assume an English-to-Chinese directionality of translation for our corpus. However, we are interested in observing how Chinese features are useful in English verb classification – our classification task is in fact uni-directional (Chinese to English). That is, we use Chinese features to aid the classification of English verbs. As it will become clear later in this chapter, the way we select our Chinese features is also uni-directional (English to Chinese). That is, within the corpus, we collect Chinese translations of a set of English verbs. Given these translations, we then look for surface features that may reveal some distinctive patterns (of different syntactic usage, for instance) for each class of English verbs.[1]

The corpus of our choice was made available by the Linguistic Data Consortium

[1]We do not deny that we are grossly simplifying the directionality issue by looking for features in one direction (English to Chinese translations) and classifying verbs in the other direction (Chinese features to English verbs). Further, there is no clean one-to-one mapping between English and Chinese lexical entries – the Chinese translations we found may be mapped to other English verbs within or outside our set of English verbs. The ideal scenario is to look for pairs of translations independent of the bitext we have. However, given the high degree of "compounding" variety of Chinese "words", it is unlikely we will find an exhaustive list of translations. We will discuss the methological issues in the next chapter.

(LDC). This corpus contains copies of the laws, press releases, and news of the Hong Kong Special Administrative Region. Because of the legal nature of the documents, this corpus is not considered a balanced corpus. However, for the same reason, we expect that the translation was done consistently (i.e., consistent many-to-many mappings between English and Chinese terms/phrases) and the two subcorpora are well aligned at the sentence level.

We performed a manual inspection of the HKLaws corpus to select a set of features for our automatic verb classification task as we describe next.

3.3 Feature Selection

3.3.1 Overview

Recall that Merlo and Stevenson (2001a) derived their features based on the linguistic distinctions among the verb classes. Our selection process is similar. The semantic distinctions among the English verb classes should be reflected in the syntax in Chinese as well. Such distinctions should also give rise to distributional differences in the syntactic behaviours. We thus need to determine such syntactic behaviours for which the amount of usage might be useful in discriminating the classes.

We followed a series of steps in the feature selection process: manual translation extraction, candidate feature selection, manual feature extraction, and final feature selection. First, both the English and the Chinese subcorpora of the HKLaws corpus were automatically POS tagged before we manually extracted a set of Chinese target verbs. Note that Merlo and Stevenson (2001a) considered verbs with the same simple past and past participle forms in order to simplify the counting process; they included only the "-ed" form of the verb, "on the assumption that counts on this single verb form would approximate the distribution of the features across all forms of the verb" (Merlo and Stevenson, 2001a). Hence, in our Chinese verb extraction task, we only considered each

English target verb tagged with the simple past (VBD) or past participle (VBN) POS tag. For each English target verb encountered in the English HKLaws, say, in some sentence i, all Chinese compounds with a verb POS tag are extracted from the near-neighbourhood of sentence i (we use ± 5 sentences) in the Chinese HKLaws. The final set of Chinese verbs was manually picked from the extracted set of Chinese compounds. We did not use a machine-readable dictionary or other automatic extraction techniques. See Chapter 4 for a brief discussion of our extraction methodology.

The second step is the feature selection step. Recall that we were interested in unergative verbs (manner-of-motion verbs), unaccusative verbs (change-of-state verbs), and object-drop verbs (creation and transformation verbs). Although we were not able to find any manner-of-motion verbs in the HKLaws corpus, we wanted to select features whose statistics potentially reflect a three-way distinction between these verb classes; these features should be manifested as (surface) syntactic features for which we can collect the frequencies from a corpus. Based on these criteria, a set of features was considered as candidates. For each feature, the number of different syntactic patterns was manually counted from a random sample of 50 Chinese sentences, each containing a translation of one of two English verbs, *open* from the change-of-state class and *build* from the object-drop class. A feature was selected if its manual count revealed enough variety (2 or more distinct syntactic and/or semantic usages). The following are the Chinese features selected:

- Chinese POS tags for verbs
- Periphrastic (causative) particles
- Passive particles
- Morpheme information
  - Different types (POS) of morpheme in each verb compound

- Semantic specificity (Does the translated verb have a meaning that is not part of the original English verb semantically?)
- Average morpheme length

Note that the above Chinese features are not completely orthogonal to Merlo and Stevenson's set of English features. That is, some of them may have some commonalities. For example, both English and Chinese have passive sentence constructions. In English, a passive sentence can be detected by the construction

be-verb ... past-participle-verb

In Chinese, a passive sentence can be detected by observing the occurrence of a passive particle preceding a Chinese verb compound. Besides the Chinese passive voice feature, some of our Chinese features are related to Merlo and Stevenson's (2001a) features linguistically, as we will discuss in the relevant sections of this chapter. In light of this, we speculate that given a semantic verb class, some syntactic features are not language-dependent, which further confirms Levin's (1993) hypothesis that the semantics of a verb is related to its syntactic behaviour.

Despite the fact that the two feature sets are not orthogonal, the Chinese features we have identified potentially aid the classification of English verbs, particularly in cases where some (related) English features are imperfectly extracted from the corpus. Moreover, the Chinese features alone may be useful in classification. We hypothesize that these Chinese features exhibit distributional differences in their frequencies of usage of the verbs. Ultimately, such differences are what is important in our automatic learning experiments. For each class of verbs, the distribution of the frequencies of feature usage can be thought of as the signature of the verb class as a whole. A machine-learning algorithm can identify the signature and therefore classify the individual verbs accordingly.

In the following sections, we will describe the selected features and their expected behaviour. Specifically, we will discuss the linguistic motivation behind the selection

of each feature, and hence, we hope that the distinguishing linguistic behaviour for the three verb classes will become clear to the reader. The linguistic analysis of each feature should shed light on its frequency patterns of usage across the classes.

3.3.2 Chinese POS tags for Verbs

The Chinese portion of the HKLaws corpus was tagged using a POS tagger provided by Academia Sinica in Taiwan. Although there are two Chinese POS tagging guidelines (one provided by the Chinese Knowledge Information Processing Group (CKIP) at Academia Sinica in Taiwan, the other by the Chinese Treebank Project at University of Pennsylvania (UPenn)), the CKIP tagger was the only Chinese POS tagger initially available. (After the data analysis was completed, the author discovered at least two other Chinese POS taggers (e.g., Lua, 1997; Zhang and Sheng, 1997), using completely different POS tags.)

Unlike the English POS tagset, in which only the tense of a verb is identified, the CKIP group claims that their annotation guidelines deal with thematic role information as well as syntactic information (Huang et al., 2000). Since our verb classes are distinguished by the different thematic roles they assign, the CKIP tags would potentially be very useful. However, not all CKIP verb tags describe the thematic role assignment of the participants involved in the state/action described by the verb. According to earlier papers describing the 15 CKIP verb tags (Liu et al., 1995; Chen and Hong, 1996), each tag describes the verb type (activity or state), and the subcategorization frame of the tagged verb. For example, a VG-tagged verb is an action verb that takes one noun phrase as a complement, but there is no indication of what thematic roles the subject and the object take. Consider the verb 磨成 (*grind*), which is given the VG tag by the CKIP POS tagger. The following two sentences illustrate the compound's alternations. Note that neither of the noun tags, *Nh* and *Na*, indicates the thematic roles for the causative agent, 他們 (*they*), the theme, 栗子 (*chestnuts*), and the oblique, 栗粉

(*chestnut-starch*).

1. (a) 他們(Nh) 把(P) 栗子(Na) 磨成(VG) 栗(b) 粉(Na) 。
they make (particle) chestnuts grind chestnut-starch
They grind the chestnuts into powder.

   (b) 栗子(Na) 磨成(VG) 栗(b) 粉(Na) 。
chestnuts grind chestnut-starch
The chestnuts grind into powder.

As shown in the above example, the CKIP group does not always consider thematic role and subcategorization frame information. However, the CKIP tags could still be very useful since the CKIP group do consider whether a verb is adjectival/stative or not. As noted by Chen and Hong (1996) and Xia (1999), all Chinese verbs, excluding the copular verbs, can be classified as either stative or non-stative. The main criterion for tagging a verb as stative is that it can be used as an adjective by appending an adjectival particle *de* (的) to the end of the verb. (However, as noted by Xia (1999), it is still an open question as to whether Chinese adjectives form a subclass of Chinese verbs.)

The state/action distinction indicated by the CKIP POS tags is potentially useful in our verb classification task. The Chinese stative verb behaviour is not unlike the adjectival behaviour observed in unaccusative verbs, but not in the other classes of verbs. For example, we can say "a *cracked* golden bowl" or "*diminished* value", but it is awkward to say "a *built* house" or "a *run* horse". Note that the **final state** of an action is part of the meaning of each unaccusative verb. This stative property makes it legal to use unaccusative verbs adjectivally. (See Section 3.3.5.1 for further discussion.)

On closer examination, the 15 CKIP verb tags can be divided into two disjoint sets, those for stative verbs and those for non-stative verbs. In other words, the CKIP tags have a finer-grained categorization of Chinese verbs than the UPenn tags. Both the CKIP and UPenn POS tagsets for verbs are used in the experiments here because it is not clear which level of categorization is appropriate.

The adjectival property of unaccusatives should lead to a higher probability of being considered stative by the POS tagger. Under the assumption that the stative/adjectival property is similar across the two languages, we expect to see that unaccusative verbs are assigned the stative verb POS tag more often than the other two classes of verbs are.

3.3.3 (External) Periphrastic/Causative Particles

As mentioned earlier, English verbs are not always translated into one single Chinese verb compound with exactly the same meaning. In some English verbs, the causative meaning is encoded in the verb. Consider the verb *crack* in the following example:

2. I cracked the egg.

For the causative construction in sentence 2, the Chinese translation could be:

3. (a) 我 打爛 蛋。
   I cracked egg

   (b) 我 將 蛋 打爛。
   I made (particle) egg crack

The second case is an instance of the use of the periphrastic particle *jiang* (將), called a "BA-construction", where *BA* is one of the most frequently used periphrastic particles (e.g., Thompson, 1973; Wu, 1996). The function of a periphrastic particle is to assign the subject as the causal agent in the event specified by the verb. We believe that, in a parallel corpus, if a verb is used causatively in one sentence of a subcorpus, then the causative meaning should be reflected in the translated sentence.

Now let's focus on the issue of causativity. According to Merlo and Stevenson (2001a), unaccusative verbs are more frequently used causatively than unergative verbs and object drop verbs. This prediction was made according to the different linguistic properties, such as the linguistic markedness, of the transitive construction across the classes. We illustrate their differences by returning to our examples in Chapter 2:

Unergative: 4. (a) The horse raced past the barn.

(b) The jockey raced the horse past the barn. (causative construction with an agentive object, *the horse*)

Unaccusative: 5. (a) The chocolate bar melted.

(b) Mary melted the chocolate bar. (causative construction with a non-agentive object, *the chocolate bar*)

Object-Drop: 6. (a) The boy played.

(b) The boy played soccer. (non-causative construction)

First, let's compare unergative verbs and unaccusative verbs in their corresponding causative constructions. Observe that unergatives undergo a restricted form of causativization in which the object is agentive. The thematic role of the agent object is "subordinated to the agent of causation", hence, this type of causativization is a rare phenomenon and is not widely attested across languages (Merlo and Stevenson, 2001a). Therefore, unergative verbs are considered rare in the causative transitive. The comparison between unaccusative verbs and object-drop verbs is more trivial. Unaccusative verbs are also used causatively more often than object-drop verbs simply because object-drop verbs are not causative verbs. On the basis of these observations, the causative feature distinguishes between unaccusatives and the other two classes of verbs in English.

Returning to the Chinese causative feature, we suggest that the English causative use is reflected in the translation. We observe that the English causative feature can provide a two-way distinction between unaccusative verbs and the other two classes. Thus, we expect the same for the Chinese causative feature as well, i.e., a more frequent use of Chinese causative particles in the translation equivalent of English unaccusative verbs.

3.3.4 (External) Passive Particles

In English, a passive sentence can be detected by observing the verb morphology and passive sentence structure. In Chinese, a passive sentence can be detected by observing a passive particle preceding the main verb. If the English passive feature is indeed useful for classifying verbs, then the relative frequency of passive particles in Chinese should be useful as well.

Merlo and Stevenson (2001a) had mixed results concerning the passive feature. The passive feature was added to their feature set as a related feature to the transitive feature, since a passive use is a transitive use of the verb.[2] The transitive feature was found to show a three-way distinction among the verb classes ($Freq_{(trans,unerg)} < Freq_{(trans,unacc)} < Freq_{(trans,objdrop)}$), and the passive feature was expected to behave similarly. Although the passive feature was found to have a positive correlation with the transitive feature ($N = 59$, $R = .36$, $p = .05$), it did not distinguish between unaccusative verbs and object-drop verbs. One possible explanation is that the passive feature is a noisy feature in English (due to tagging errors and the lack of overt evidence for the passive in some English constructions).

The motivation for exploring the passive feature in Chinese is that it is a more easily detectable feature, which could lead to counts which are more useful than the English passive counts. Given the linguistic relation between the passive and the transitive, and the three-way distinction made by the English transitive feature, we hypothesize that the (overt) Chinese passive feature will also make a three-way distinction among our verb classes.

[2]For example, according to Trask (1993), a passive construction is "[a] construction in which an intrinsically transitive verb is construed in such a way that its underlying object appears as its surface subject."

3.3.5 Morpheme Types

The extracted target verbs in Chinese consist of a varying number of morphemes. This is partly because some English verbs may not be translated into one simple verb compound with exactly the same meaning. If an English verb has a translation equivalent in Chinese, the translated verb compound tends to have no more than two morphemes (sometimes with one of the morphemes encoding the complement of the English verb). However, some English verbs may require more morphemes to encode a similar meaning. In some cases, serial verbs (a string of two or more morphemes in serial) are required.

Note that the verbs considered here are manner-of-motion, change-of-state, creation, and transformation verbs. The question here is, is the semantic class membership of an English verb related to the types of sublexical elements used in the Chinese equivalent? If so, do the individual morphemes of a Chinese verb compound reflect the semantic class membership? The following three features are attempts to address this issue.

3.3.5.1 Compounding Pattern

Chinese compounds in general have a high variety of compounding patterns. Many permutations of morphemes with different parts-of-speech are legal. We observe that there might be a relationship between the semantic property of a compound and its morphology. This notion that a compound's (semantic) class is related to the types of morphemes in the compound is not new. For example, Hsu and Wu (1994) examined many possible Chinese translations of the verb *break* and they concluded that change-of-state verbs are more likely to be translated as V-A compounds, where the second morpheme indicates the final state as a result of the action indicated by the first morpheme. Many Chinese examples of change-of-state verbs in the form of V-V or V-A combination are listed in (Chang, 1990; Lin et al., 1997; Chang, 1998; Starosta et al., 1998). However, based on Huang's (1998) study on the headedness of Chinese disyllabic compounds, all permutations are possible for verbs. Hence, we want to examine all possible patterns in the

translation, not just V-V or V-A combinations.

To follow Hsu and Wu's (1994) idea that English change-of-state verbs can be translated into Chinese resultative verb compounds, let's examine resultative constructions more closely. According to Levin and Rappaport Hovav (1995), a resultative construction indicates a change of state where both the activity and the result state are lexically specified, either in the same verb or in separate phrases. For example, in the following sentence, a change-of-state verb specifies both the activity and the final state.

7. Dead skin cells clogged the pores on my face.

The verb *clog* specifies both the *clogging* activity and the *clogged* result state of the pores. Note that the verb in a resultative construction does not always indicate the change of state when used alone. Consider the following sentences using the verb *run*:

8. (a) The joggers ran.

   (b) The joggers ran the pavement thin. (Levin, 1993)

Here in the first sentence, the verb *ran* alone does not indicate a change of state, but in the second sentence, the pavement became thin as a result of the joggers running. Unlike sentence 7 in which the resultative meaning is obligatory, the final state of the pavement must be explicitly stated in sentence 8b to be considered resultative.

We see that both manner-of-motion and change-of-state verbs can be in resultative constructions; we must also ask if object-drop verbs participate in resultative sentences. Levin and Rappaport Hovav (1995) suggest that interpreting a sentence using a creation/accomplishment verb as a resultative sentence can be problematic. Consider this example,

9. We built the house.

Only the result state *the house is built* but not the change-of-state sub-event (i.e., what are the "before" and "after" states of the house) is specified by the verb. In sentence 7,

we saw that *the pores* went through a change-of-state sub-event from being unclogged to being clogged. Similarly, *the pavement* went through a change-of-state sub-event from being "thick" to being "thin". However, we cannot interpret sentence 9 the same way. *The house* was non-existent before the *building* event. It did not go through a change of state.

Based on the above line of reasoning, we hypothesize that there is a potential three-way distributional differences in the amount of resultative constructions among the verbs. Unaccusative verbs should have the most frequent resultative use because of the change-of-state meaning. Object-drop verbs should have the least resultative use because of **the lack of** the change-of-state meaning. The frequency of resultative use of unergatives should be in between those of the unaccusatives and the object-drops because the change-of-state meaning can be added optionally.

We suggest that the change-of-state meaning of English resultative constructions is reflected in the translation as suggested by Hsu and Wu (1994). In particular, the change-of-state meaning is more likely to be manifested as a V-V or a V-A combination because both the action and the final state are specified in the morphemes. Therefore, we hypothesize that English unaccusative verbs are the most likely to be translated into Chinese V-V or V-A verbs, followed by unergative verbs, and then object-drop verbs.[3]

3.3.5.2 Semantic Specificity

We now consider a slightly different type of information encoded in the individual morpheme. Specifically, some Chinese verbs are "semantically more specific" than the English

[3]In (Liu et al., 1995), the CKIP tagset included one tag, *VR*, for resultative verbs. However, the authors included rules for tagging all verb tags except *VR*. In a subsequent paper (Chen and Hong, 1996), the authors discarded the *VR* tag altogether. If there were a POS tagger which made use of this tag, the detection of Chinese resultative verbs would be easier.

Another way of detecting resultative constructions is to count the number of collocations of a verb and the morpheme 了. This morpheme can be used as an aspectual marker which sometimes indicates an endpoint. However, the difference between grammatical aspect (an event happens in the past) and lexical aspect (an event which has an endpoint) is ambiguous.

equivalent. For example, 改建 encodes more information than the verb *alter* alone. It carries the connotation of *rebuild* as well. However, both 遊戲 and 放映 simply give two different senses of the same verb *play*. The first one means *play* as in *play a game*; the second is *play* as in *play a movie*. Nothing extra is added to the meaning.

Both Hsu and Wu (1994) and Palmer and Wu (1995) examined many possible translations for the change-of-state verb *break*. They observed that one single lexical entry *break* corresponds to many Chinese verb compounds representing a wide range of breaking events. Each compound includes information about both the activity and the final state. For example, 夾碎 means *clamp-into-pieces* where 夾 corresponds to the clamping action and 碎 corresponds to the final *broken* state. From the previous section, we know that a Chinese resultative verb encodes two pieces of information: the activity and the final state as a result of the activity. Either morpheme can encode more information than the meaning of the original English verb alone.[4]

We speculated earlier that there is a higher frequency of the use of English unergative and unaccusative verbs in resultative constructions. If this assumption holds, then a higher frequency of resultative constructions should lead to a higher frequency of lexicalizing the activity and the resultative state in the individual morphemes of verb compounds, and hence a higher frequency of "semantically more specific" Chinese translations for unergative and unaccusative verbs.

3.3.5.3 Average Morpheme Length

Counting the number of morphemes is yet another way of looking at morpheme encoding of meaning. For example, we may need more morphemes to encode the two pieces of information needed in a resultative construction. Let's return to our example in

[4]We mentioned that in English, both the activity and the final state are implied in a change-of-state verb. For example, both the breaking event and the broken final state are implicit in the verb *break* (and other *break* verbs such as *smash*). The point here is that Chinese change-of-state verbs usually require two separate morphemes to encode the core event and the final state, with at least one of them more **specific** and **explicit** than the English meaning.

Section 3.3.5.1,

10. The joggers ran the pavement thin. (Levin, 1993)

One possible translation is:

11. 運動員 把 行人道 跑薄 了。
    The joggers make (particle) pavement run thin le (resultative particle)

In this example, there are two morphemes in the Chinese verb compound 跑薄 in which 跑 is the running action and 薄 describes the final thin state of the pavement. Observe that the first morpheme 跑 alone is a translation of *run*. Typically, extra morphemes are optional for the Chinese equivalent of *run*. However, in this example, the second morpheme is necessary to indicate the resultative state. We have previously noted that resultative constructions describe complex events with multiple sub-events (the action that causes a change-of-state and the end state of the action) while non-resultative sentences describe only the activity. We postulate that "complexity" of an event and the number of morphemes needed to encode it are correlated.

As mentioned in earlier sections, unergatives and unaccusatives are used in resultative constructions more frequently. Furthermore, it is optional to use unergative verbs resultatively. Generally, we expect Chinese translations of unaccusative verbs to have the highest average morpheme length, followed by unergative verbs, and then object-drop verbs.

3.3.6 Summary of Features and Their Predicted Behaviour

As seen in the previous sections, we have selected six Chinese features. By analyzing the translations of our target English verbs, we derived Chinese features that are correlated with English verb behaviour either syntactically (i.e., passive voice construction) or semantically (i.e., adjectival behaviour, causativity, and resultative construction). Table 3.1 gives a summary of the Chinese features and the different expected frequency

patterns depending on the (English) verb class membership. In the next chapter, we will describe how we collect or approximate the frequency counts for these features.

Chinese Feature	Expected Frequency Pattern	Explanation
(Stative) POS tag	Unerg, ObjDrop < Unacc	Unaccusative verbs can be adjectivized and hence are more likely to be translated into Chinese stative verbs.
Periphrastic particles	Unerg, ObjDrop < Unacc	The use of the periphrastic particle is correlated with the English causative feature.
Passive particles	Unerg < Unacc < ObjDrop	The use of the passive particle is correlated with the English passive voice feature.
Morpheme patterns: resultative constructions	ObjDrop < Unerg < Unacc	Only unaccusative and unergative verbs can be used resultatively. The resultative meaning of unaccusatives is obligatory, but it is not the case for unergatives. Hence, unaccusative verbs have the highest likelihood of being translated into Chinese resultative verbs, followed by unergative verbs, and then object-drop verbs.
Morpheme patterns: semantic specificity	ObjDrop < Unerg < Unacc	This feature is correlated with the resultative construction.
Morpheme patterns: average morpheme length	ObjDrop < Unerg < Unacc	This feature is correlated with the resultative construction.

Table 3.1: The Chinese features and their expected behaviour.

Chapter 4

Data Collection

In the previous chapter, we described the set of Chinese features we intend to collect statistics on. Since our work is based on (Merlo and Stevenson, 2001a), we also collect statistics for their set of English features. However, in this chapter, we will focus primarily on how we collect the frequencies of the Chinese features.

4.1 Materials and Method

In Chapter 3, we described a feature selection process using the HKLaws corpus. The same corpus is used as our primary source of data. Recall that this corpus is a sentence-aligned bilingual corpus with approximately 6.5 millions words in the English subcorpus and 9 million characters in the Chinese portion. Note that the English HKLaws is about 10% of the combined size of years 1987-1989 of the *Wall Street Journal* (WSJ) and the Brown Corpus used by Merlo and Stevenson (2001a). To augment the data extracted from HKLaws, the WSJ is used as an additional source of English data.

4.1.1 Manual Corpus Analysis

As mentioned in section 3.3, a manual analysis step precedes the feature selection step. For each English target verb that appears in the English HKLaws, the corresponding Chinese verb is manually extracted. The manual extraction is necessary because many combinations of morpheme stems are possible (Huang, 1998). One cannot expect a dictionary (machine-readable or otherwise) to contain all possible English-to-Chinese translations. Hence, using a machine-readable English-to-Chinese dictionary is not a feasible way to extract all the Chinese verbs in the corpus accurately.

Due to the legal nature of the HKLaws corpus, we could not find any unergative (manner-of-motion) verbs. Our list consists of 16 unaccusative (change-of-state) and 16 object-drop (creation and transformation[1]) verbs. Some of the Chinese verbs may have extra meaning that is not a necessary component of the original English verbs, but they are, in the author's opinion, all possible translations of the English target verbs. See Tables 4.1 and 4.2 for the verbs we used in our experiments.

One possible way to automate the extraction process would be to use a type of bitext alignment or lexicon extraction algorithm (e.g., Fung and McKeown, 1997; Melamed and Marcus, 1998; Pao, 2000). Clearly, such a fully automated extraction method would introduce noise into the data. Given the relatively small size of the HKLaws corpus, such noise may obscure the patterns in the accurate data. Recall that our goal is to explore whether bilingual information can improve fine-grained verb class distinction. Developing a fully automatic technique for bilingual verb class acquisition is outside the scope of our investigation.

[1]The only exception is the verb *pack*. Based on (Levin, 1993), the verb *pack* is neither a creation verb nor a transformation verb. However, it is still an object-drop verb. The reason we included it is that we could not find another creation or transformation verb in the HKLaws corpus. In order to have an equal number of unaccusative verbs and object-drop verbs, we selected *pack*, which is included in (Merlo and Stevenson, 2001a), as one of the object-drop verbs we considered.

English Verbs	Chinese Verbs
alter	更改, 修改, 改動, 改建, 竄改, 改裝, 改變, 變更
change	轉變, 轉換, 改變, 修改, 更改, 更換, 修改, 變化, 變動, 變更, 改爲, 改名爲, 變, 改
clear	剔除, 清除, 結算, 清除, 清理
close	結算, 密封, 閉封, 密閉, 關閉, 圍封, 閉合, 封閉, 封蓋, 完成, 停止, 終結, 停開, 關緊, 結束
compress	壓縮
contract	招致, 訂約, 締結, 約定, 簽訂, 立約, 訂立, 訂有, 訂明, 所患, 患染, 罹患, 患上
cool	冷卻, 散熱, 冷
decrease	降低, 調低, 減低, 扣除, 增減, 減少
diminish	消損, 貶值, 減少, 縮減, 減去, 刪去, 減損, 減責, 調減, 降低, 減值
dissolve	解散, 解除, 溶解
divide	分配, 劃分, 分成, 分給, 分爲, 分拆, 拆分, 分割, 除以, 分開, 分間, 相等
drain	排水, 排去, 排放, 滴乾, 排出
flood	浸水, 浸, 淹沒
multiply	乘以, 乘, 倍
open	開放, 開展, 開啓, 打開, 破啓, 公開, 開設, 開立, 開掘, 掘開, 揭開, 開膛, 張開, 開始, 拆開, 開
reproduce	複印, 重現, 複製, 重播, 生殖, 載於, 轉錄, 列出, 反映, 複載, 重複

Table 4.1: English unaccusative verbs and their corresponding Chinese verbs extracted from the HKLaws.

4.1.2 Chinese Feature Extraction

An English verb can be translated into one or more Chinese verb compounds, and vice versa. That is, there is a many-to-many mapping from the set of English verbs to the set of Chinese verbs. To simplify our task, we only look at the one-to-many English-to-Chinese mapping. That is, for each English verb, e_i, we have a vector of features, $f_1, \ldots, f_p, f_{p+1}, \ldots, f_m$, where a subset of these features, $f_{p+1}, \ldots, f_m$, are Chinese features. Let $C_i = c_{i1}, \ldots, c_{in}$ be the set of n possible Chinese translations for the verb e_i. For each Chinese verb, c_{ij}, $j = 1 \ldots n$, we count the number of occurrences of each feature, f_k, $k = p+1, \ldots, m$, in the corpus. The sum of these frequencies is the final raw frequency of the feature, f_k. The raw frequency is then normalized.

English Verbs	Chinese Verbs
build	裝置, 製成, 安裝, 建造, 建成, 興建, 建
clean	清潔, 清理, 潔淨
compose	組成
direct	作出, 致予, 指示
hammer	鎚擊
knit	針織, 紡織, 織
organise	成立, 組織
pack	包裝, 裝載, 裝入
paint	繪畫, 油漆, 漆上, 畫上, 髹
perform	辦理, 履行, 進行, 執行, 實行, 舉行, 表演
play	遊戲, 放映, 展示, 播放
produce	出示, 交出, 交給, 製作, 提供, 呈交, 提交, 產生, 生產, 製成
recite	列舉, 敘述, 詳述, 述
stitch	縫製, 縫合
type	打字
wash	清洗, 洗淨, 洗濯, 洗

Table 4.2: English object-drop verbs and their corresponding Chinese verbs extracted from the HKLaws.

Given this basic feature extraction method, we refine it in two different ways. We refer to them as the "aligned" and the "unaligned" methods. The first method is the "aligned" method: for each English target verb encountered, we consider the Chinese verb in the corresponding aligned sentence. Although the HKLaws corpus is considered a sentence-aligned corpus, in reality the two sub-corpora are not perfectly aligned. In some instances, the supposedly aligned sentence pair is offset by 10 or more sentences. When a Chinese verb is not found in the supposed target location, we use a window of ± 30 sentences. To avoid double-counting a Chinese verb in previously visited sentence indices, we keep track of a list of seen sentences. As long as we do not overcount, we consider all Chinese target verbs within this window. Obviously, there are cases where an English sentence is not translated to a Chinese sentence with a nice one-to-one part-of-speech mapping, i.e., no Chinese target verbs can be found in the target window. As a result, for each English target verb, the number of Chinese verbs found does not always

match the English frequency.

The other extraction method is the "unaligned" method. Unlike the first method, this method does not rely on the sentence alignment in a parallel corpus. Given the same set of Chinese translations used in the aligned method, we consider all target Chinese verbs encountered as we process the corpus sentence by sentence.

Note that in either method, these frequencies are relative frequencies. Except where otherwise noted, for each Chinese feature of each English verb, the relative frequency is obtained by using the number of all Chinese translations encountered as the divisor.

4.1.2.1 Chinese POS tags

The count for each CKIP verb tag is simple. For each verb of interest, we collect the number of occurrences of each POS label in the tagged corpus. To get the equivalent count for the UPenn tags, each CKIP tag is simply mapped to either VA (stative) or VV (non-stative) (Fei Xia, personal communication).

4.1.2.2 (External) Causative/Periphrastic Particles

If a periphrastic particle precedes a Chinese target verb, this is counted as one occurrence of the particle. We only consider a particle which precedes the target verb in the same clause. For example,

... punctuation ... particle ... target verb ... punctuation ...

Although there are legal cases where there is punctuation between the particle and the verb, for example,

... particle ... punctuation ... target verb ...

they do not happen very often. We ignore these cases in order to avoid counting those particles that are not related to the target verb.

4.1.2.3 (External) Passive Particles

The method of counting passive particles is similar to the one for periphrastic particles. For each verb, if a passive particle precedes a target verb, and both occur between two punctuation marks, this is counted as one occurrence of the particle.

Note that syntactically, a passive Chinese sentence appears exactly the same as a periphrastic Chinese sentence (NP0 + Particle + NP1 + VP). The CKIP POS tagset does not distinguish the type of particle used. To distinguish the two, we can perform simple pattern matching by keeping a hash table of all passive and periphrastic particles.

4.1.2.4 Sublexical Information

Compounding Pattern The following morpheme patterns are possible in a Chinese verb compound: adj-adj (A-A), adj-noun (A-N), adj-verb (A-V), noun-adj (N-A), noun-noun (N-N), noun-verb (N-V), verb-adj (V-A), verb-noun (V-N), and verb-verb (V-V). Note that each morpheme can belong to multiple parts-of-speech. Since most contemporary Chinese dictionaries do not list the part-of-speech of individual characters, we manually compiled a list of morpheme–part-of-speech correspondences. To find the part-of-speech of a morpheme, we simply match it against the list.

Consider the compound 破啓 as a translation of the verb *open*. The first morpheme 破 can be either a verb (*to break*) or an adjective (*broken*). The second morpheme is 啓 which is a verb (*to open*). For this particular example, we have two possible combinations: V-V and A-V. Counts for both V-V and A-V are incremented. Since the number of possible combinations can exceed the number of translations, we take the sum of the raw counts for each pattern as the divisor.

Semantic Specificity For each Chinese target verb encountered, if it is manually annotated as "semantically more specific" (than the original English verb) as described in section 3.3.5.2, this is counted as one occurrence of this feature.

Average Compound Length We define the compound length as the number of morphemes in a Chinese verb compound. To collect the statistics for this feature, we can sum the number of morphemes in each Chinese target verb encountered and take the average.

4.1.3 English Feature Extraction

The set of English features is exactly the same as the set used by Merlo and Stevenson (2001a). Recall that the features are animacy, causativity, passive voice, transitivity, and VBN POS tag. In (Merlo and Stevenson, 2001a), the features transitivity, passive voice, and VBN were extracted from the POS tagged version of the WSJ available from the LDC. The remaining two features, causitivity and animacy, were extracted from a parsed version of the 1989 WSJ. Our extraction technique is almost exactly the same except that no parser or chunker was used. Instead, the count for each feature is extracted by means of a set of regular expressions, provided by Merlo and Stevenson, which were applied to the POS tagged English HKLaws. Because of the relatively low frequency of our verbs in the corpus, we used the WSJ (as in (Merlo and Stevenson, 2001a)) as an additional data source. The English HKLaws corpus was automatically tagged using Ratnaparkhi's maximum entropy tagger (Ratnaparkhi, 1996). The POS tagged version of the WSJ was provided by the LDC. (Merlo and Stevenson (2001a) suggested that "a fully parsed corpus is not necessary ... a more accurate tagger ... might be sufficient to overcome the fact that no full parse is available." In fact, their results were replicated in English without the use of a parser by Anoop Sakar and Wootiporn Tripasai (Anoop Sakar, personal communication).)

Note that the notion of aligned and unaligned frequencies mentioned in the previous section does not apply here; that is, there is only one set of English data per corpus. This is because the English features were extracted independent of the Chinese data. The English data obtained here serve two purposes: (i) to replicate the experiments

of Merlo and Stevenson (2001a); (ii) to extend Merlo and Stevenson's experiments by combining the English data with the Chinese data. (The pairing of English and Chinese data in the experiments will be discussed in the next two chapters.)

In the original data collection plan, we used a strict clause-based extraction method. That is, for each English verb, we collected all occurrences of the "-ed" form with a verb POS tag with the restriction that the target verb occurs within a clause that is at the beginning of a sentence or at the beginning of a clause indicated by a punctuation. Although these search patterns are merely regular expressions, we considered them as a limited set of grammatically correct clausal patterns. However, the number of verbs found in HKLaws using this method is in the single digit range. In the hope of finding more occurrences, we relaxed the regular expressions by removing the "start of a sentence/clause" restriction from the search pattern, although using the relaxed regular expressions is problematic (see our analysis in section 4.2.2).

4.2 Data Analysis

Using the relaxed search patterns, we were able to collect the following English data points from the HKLaws corpus: transitivity: 101; passive voice: 2,468; VBN: 2,954; causativity: 443; animacy: 28. Since we have 32 verbs in total, these raw frequencies are by no means high. Using the WSJ, we were able to obtain more data without using any relaxed search patterns. The data points are: transitivity: 21,054; passive voice: 5,376; VBN: 23,238; causativity: 15,351; animacy: 1,303.

For the Chinese data, we had two extraction methods as described in Section 4.1.2. The aligned method yields these data points: Chinese verb POS tags: 4,100; passive particles: 52; periphrastic particles: 63; morpheme patterns: 5,904; semantic specificity: 119. The unaligned method yields these data points: Chinese verb POS tags: 113,087; passive particles: 453; periphrastic particles: 1,039; morpheme patterns: 113,617; seman-

	Unaccusative				Object-Drop			
	alter		*flood*		*play*		*wash*	
Feature	Man	Auto	Man	Auto	Man	Auto	Man	Auto
CKIP VB Tag	0	0	1	1	0	0	0.02	0.02
CKIP VC Tag	1	1	0	0	1	1	0	0
CKIP VCL Tag	0	0	0	0	0	0	0.98	0.98
Pass.Part.	0	0.06	0.13	0.13	0	0	0	0
Peri.Part.	0.08	0.03	0	0	0	0	0	0
A-V Morph.	0	0	0	0	0	0	0.40	0.40
V-A Morph.	0	0	0	0	0	0	0.12	0.12
V-N Morph.	0	0.04	1	1	0	0	0	0
V-V Morph.	1	0.96	0	0	1	1	0.48	0.48
Sem.Spec.	0	0.05	0	0	0	0	0	0
Avg.Length	2	2	1.78	1.78	2	2	1.93	1.93

Table 4.3: Manually and automatically calculated Chinese feature frequencies of a random sample of verbs, aligned method

tic specificity: 4,063.

4.2.1 Chinese Data

To ensure the collected data indeed reflects the lexical and sublexical properties as described in section 3.3, we compare our manual sampling counts (see sections 3.3 and 4.1.1) with the automatic counts. Specifically, we picked two verbs from each class, 50 sentences per Chinese verb (or the total number of Chinese verb instances, whichever is less), and manually counted each feature. The four verbs we did a hand-count on are: *alter* (aligned frequency: 312; unaligned frequency: 4,158), *flood* (aligned frequency: 23; unaligned frequency: 197), *play* (aligned frequency: 10; unaligned frequency: 1,114), and *wash* (aligned frequency: 43; unaligned frequency: 328).

Table 4.3 shows the manually and automatically counted Chinese feature frequencies of the first 50 occurrences or the total number of occurrences, whichever is less, of the Chinese verbs using the aligned method. (Only the rows with non-zero entries are shown. For example, there are a total of 15 verb POS tags and 9 morpheme combinations.

	Unaccusative				Object-Drop			
	alter		*flood*		*play*		*wash*	
Feature	Man	Auto	Man	Auto	Man	Auto	Man	Auto
CKIP VA Tag	0	0	0	0	0	0	0.06	0.082
CKIP VB Tag	0	0	0.90	0.94	0	0	0	0
CKIP VC Tag	1	1	0.10	0.06	1	1	0.94	0.91
CKIP VH Tag	0	0	0	0	0	0	0	0.01
Pass.Part.	0	0.06	0.14	0.13	0	0	0	0
Peri.Part.	0	0.02	0	0.01	0	0.02	0	0.02
A-A Morph.	0	0.01	0.19	0.10	0	0	0	0
A-V Morph.	0	0	0	0	0	0	0.39	0.40
V-A Morph.	0	0	0.19	0.10	0	0	0.04	0.02
V-N Morph.	0	0.04	0.63	0.80	0	0	0	0
V-V Morph.	1	0.97	0	0	1	1	0.57	0.58
Sem.Spec.	0	0.05	0	0	0	0	0	0
Avg.Length	2	2	1.44	1.51	2	2	1.68	1.70

Table 4.4: Manually and automatically calculated Chinese feature frequencies of a random sample of verbs, unaligned method

Most verbs are tagged with at most three or four different POS tags and morpheme combinations.) Most of the frequency pairs are exactly the same or extremely close. The few discrepancies come from the verb *alter*. The reason is that some occurrences for passive particles, V-N morpheme combinations, and semantically more specific verbs do not show up until much later in the corpus and they are few in number. Despite this, we consider the automatic counts in the aligned method accurate.

We also conducted a hand count using the unaligned method. Our manual corpus analysis revealed that given an English verb, the translation is based on the topic of a particular section. If we picked only the first 50 occurrences, the counts would likely be biased towards the type of translation in the earlier sections. What we did instead was to pick a random sample of 50 sentences per verb. The counts are shown in Table 4.4. Again, only the non-zero rows are shown.

In comparison to the aligned counts, the unaligned counts have more discrepancies between the manual calculated frequencies and the automatically calculated frequencies.

	Unaccusative				Object-Drop			
	alter		*flood*		*play*		*wash*	
Feature	Man	Auto	Man	Auto	Man	Auto	Man	Auto
Causativity	0	0.014	0	0.25	0	0	0	0
VBN	1	0.979	1	1	0.59	1	0.96	0.803
Passive Voice	0.96	0.935	1	1	1	0.857	0.96	0.824
Transitivity	0.56	0.883	0.79	0.917	1	0.889	0.96	0.824
Animacy	0	0.118	0	0	0	0	0	0

Table 4.5: Manually and automatically counted English feature frequencies for a random sample of verbs extracted from the HKLaws

However, the differences are very small. We speculate that using a small sample (50 random instances per verb) results in a much coarser set of hand counts. In general, we consider the counting procedure sufficiently accurate for our purpose.

4.2.2 English Data

As mentioned earlier, we relaxed the English search pattern to allow mid-sentence matches. However, using the relaxed patterns can be problematic. For each of the same four verbs, we also performed a hand count for each feature in the first 50 occurrences (or all occurrences, whichever is smaller). Let us look at a comparison of the manual count and the automatic count in Table 4.5. As we can see, there are considerable differences between some of the manual and automatic counts. Given these differences, we believe that the English HKLaws data may not be very useful for our task for the following reasons:

Limited Search Patterns The search patterns do not cover all the complex cases. For example, the patterns can handle cases such as

> ... *where engagement is* cancelled or *altered*.

but not

> ... *makes use of any licence that has been forged or*, without the authority

of the director, *altered*.

These patterns cannot handle cases where there are long phrases cutting in between the auxiliary verb (in the above examples, *is* and *has been*) and the main verb (*altered*).

Errors Caused by "Relaxing" the Search Patterns Relaxing the patterns to allow mid-sentence matches can be problematic. Consider the following POS tagged clause.

> `The/DT person/NN to/TO whom/WP the/DT first/RB mentioned/VBN person/NN is/VBZ ,/, by/IN` *reason/NN of/IN the/DT form/NN of/IN marriage/NN contracted/VBD* `,/, lawfully/RB married/VBN` ...

The italicized portion is matched by the search pattern for an active clause with the subject *reason* ("modified" by two prepositional phrases, *of the form* and *of marriage*) and verb *contracted*. If we considered only clauses at the beginning of a sentence or immediately after a punctuation, these mismatches would not happen at all. However, by following this strict sentence/clause-beginning restriction, we were able to match very few sentences. For this particular example, we do not have patterns dealing with the middle portion *"by reason of the form of marriage contracted"*. This is why we relaxed the search pattern in the first place.

Errors from the POS tagger The inaccuracy of the POS tagger also causes the search patterns to fail sometimes. For example, when the main verb is far away from the auxiliary verb, the main verb is sometimes not tagged correctly. Consider this example:

> `...any/DT licence/NN that/WDT has/VBZ been/VBN forged/VBN or/CC ,/, without/IN the/DT authority/NN of/IN the/DT director/NN ,/,` *altered/VBD* `,/, ...`

In this example, the verb *altered* should be tagged as VBN, not VBD. If the sentence were correctly tagged, it would be matched with a passive-voice pattern. However, as men-

	Unaccusative				Object-Drop			
	alter		*flood*		*play*		*wash*	
Feature	Man	Auto	Man	Auto	Man	Auto	Man	Auto
Causativity	0	0.06	0	0.05	0.04	0.40	0	0
VBN	0.50	0.82	0.52	0.76	0.52	0.38	0.61	0.82
Passive Voice	0.17	0.14	0.17	0.28	0.11	0.06	0.12	0.21
Transitivity	0.60	0.52	0.72	0.60	0.57	0.54	0.33	0.35
Animacy	0.16	0.15	0.08	0.05	0.18	0.21	0.33	0.17

Table 4.6: Manually and automatically counted English feature frequencies for a random sample of verbs extracted from the WSJ

tioned earlier, our search patterns are not general enough to pick up complex situations like this one.

For the data obtained from the WSJ, we did not use the relaxed search patterns. For the same four verbs, we performed a hand count for each feature for a random sample of 50 occurrences. The automatic and manual counts are shown in Table 4.6. Observe that there are some differences between some of the hand and automatic counts, even though we did not use any relaxed search patterns. We speculate that the main culprit is that we have limited search patterns to provide enough data. More (accurate) search patterns may increase the accuracy.

Chapter 5

Experimental Results

To see if we can aid the automatic verb classification task using our set of Chinese features, we will add our features to Merlo and Stevenson's vector (of English features) for each English verb, resulting in a vector of a larger dimension:

Vector Template:	[verb,	English Features,	Chinese Features,	class]
Example:	[altered,	0.04, . . . , 0.12,	0.03, . . . , 2,	2]

In order to compare our results with theirs, we will also feed the resulting set of 32 vectors into the same machine learning system C5.0 (`http://www.rulequest.com`). Recall that we have 16 unaccusative verbs and 16 object-drop verbs (see Chapter 4). The baseline accuracy is 50%; that is, chance performance would classify one of every two verbs correctly. The theoretical maximum accuracy is 100%. However, for a 3-way classification task, the best performance achieved by a group of human experts is 86.5% (Merlo and Stevenson, 2001a). Although our set of verbs is different from those used by Merlo and Stevenson (2001a), the expert-based accuracy suggests that a more realistic upper bound for our classification task would be less than the perfect accuracy.

5.1 Experimental Methodologies

In our experiments, we used both N-fold cross-validation and leave-one-out methodologies. In each iteration of an N-fold experiment, a random subset of $32/N$ vectors is used as testing data and the remaining vectors are used as training data. For each set of features, we were able to obtain an average accuracy across the folds as a performance measure. The motivation for performing N-fold cross-validation experiments was to evaluate the contribution of each feature to learning. We wanted to find the best feature combination(s). Ideally, we would like to perform an exhaustive set of $2^{21} \approx 2$ million experiments (all possible combinations of 21 features, with the 15 CKIP POS tags considered as one group). Although each classification experiment takes no more than 5 seconds of CPU time to finish, the classification experiments themselves are not fully automatic (we need to generate the feature set we want to use for an experiment, manually or by means of a shell script). Furthermore, the analysis of the classification results is not straightforward and is quite time-consuming. (In short, say, we want to compare the results using feature sets A, B, and C. For each set, we used a shell script to extract the results from a C5.0 generated file and output them to a column of a summary file. We wrote a MATLAB script to read from the summary file and perform a one-way ANOVA test. To ease the analysis, the script displays only the pairs of feature sets with results that are considered statistically distinct by MATLAB.) Therefore, finding the best combination(s) is nearly impossible. We instead attempted to assess the usefulness of each feature by observing:

- the performance using individual features
- possible performance degradation from removing individual features from a set of features
- possible performance gain by adding individual features to a set of features.

We selected the features that contribute highly in these experiments. These features were then used in the leave-one-out experiments.

The leave-one-out experiments complement the cross-validation experiments. Note that the leave-one-out experiments are similar to the cross-validation experiments with $N = 32$ because we train on $N-1$ verbs and test on the remaining verb in each iteration. However, random subsets of verbs were used in the cross-validation experiments and hence we were only able to calculate the average accuracy across the folds (or the repeated runs; see Section 5.2 for details). From the leave-one-out experiments, we have access to the classification result of individual verbs. This allows us to calculate the precision and recall for each verb class. In the following sections, we will describe our experiments in further detail and report our results.

5.2 Results Using N-Fold Cross-Validation

This section reports the results of experiments using the N-fold cross-validation training method. That is, our set of 32 vectors is randomly divided into N parts. We leave out a testing set of $32/N$ vectors and train on the remaining data. The procedure is repeated N times with different training and testing sets each time. As a result of the N runs, we have an average accuracy and a standard error. Note that our data set is even smaller than the set used by Merlo and Stevenson (2001a), so to avoid the problems of outliers, we also repeated the N-fold cross-validation procedure 50 times, as in their experiments. For each experiment, we report only the average accuracy and the average standard error over the 50 runs.

In our experiments, we picked N as 8 simply because it evenly divides 32. By inspection, the performance of 8-fold experiments show little to no difference from the 10-fold experiments. We also empirically tested the tuning options available in C5.0. Except for the tree pruning percentage, we found the remaining tuning options offer little to no

improvements over the default settings. We set the pruning factor to 30% for the best overall performance over a variety of combinations of features. (According to the manual, the default is 25%. A larger pruning factor results in less pruning in the decision tree.)

In the following subsections, we report our results for each combination of datasets we used. Since we collect each set of monolingual data independently, we have a total of four datasets: English HKLaws dataset, English WSJ dataset, aligned Chinese HKLaws dataset, and unaligned Chinese HKLaws dataset. This allows us to look at the datasets in eight different ways: the four datasets individually, and the English and Chinese datasets paired up in four different combinations. Recall that our goal here is to evaluate the contribution of each feature to learning. To do so, the set of features we used comes in several flavours: we used individual monolingual features alone, all monolingual features combined, the complete bilingual feature set with individual features removed, and the complete English feature set augmented with individual Chinese features. In particular, we want to observe the following:

- the performance of each feature alone (does this feature have an above-chance accuracy?);
- the possible performance degradation from removing a feature from the full bilingual feature set (does this feature contribute to learning?);
- the possible performance gain by adding one Chinese feature to the full English feature set (is there any performance benefit by adding this single Chinese feature to the set of English features?).

We will address the above in the next two sections: we will present results using strictly HKLaws data, followed by results using combinations of English WSJ data and Chinese HKLaws data.

5.2.1 HKLaws Data

In this section, we report results using data collected strictly from the HKLaws corpus.

Table 5.1 shows the performance of classification using individual English features. None of the English features has a better-than-chance performance.

English Feature	%Accuracy	%SE
Causativity	48.5	0.5
VBN	48.8	0.5
Passive Voice	48.6	0.5
Transitivity	49.5	0.5
Animacy	45.4	0.5

Table 5.1: Percent accuracy and standard error of the verb classification task using each English feature individually, with 8-fold cross-validation training method repeated 50 times.

Table 5.2 shows the performance of classification using individual Chinese features. The left panel of the table shows results using Chinese data collected using the aligned method, and the right panel shows results using the unaligned method. The shaded cells highlight all the better-than-chance features in this experiment. Contrary to the English results, many Chinese features show a better-than-chance performance, except the *A-X Morph.*, *N-X Morph.*, *V-V Morph.*, *Avg. Length*, for both aligned and unaligned methods, and *Sem. Spec.* for the unaligned method. All the better-than-chance accuracies are statistically distinct from one another ($p < .05$), using an ANOVA with a Tukey-Kramer post-test, except for between *CKIP Tags* and *UPenn VA-Tag*, *CKIP Tags* and *Pass. Part.*, *UPenn VA-Tag* and *UPenn VV-Tag*, *UPenn VA-Tag* and *All Verb Tags*, *All Verb Tags* and *Pass. Part.*, *Peri. Part.* and *Sem. Spec.*, and *V-A Morph.* and *V-N Morph.* (From now on, all differences in performance mentioned in this section are statistically significant at $p < .05$ using an ANOVA with a Tukey-Kramer post-test.)

Now we turn to our results using all features and all features minus one feature. Recall that our goal for these experiments is to observe the possible performance degradation

Aligned Feature	%Accuracy	%SE	Unaligned Feature	%Accuracy	%SE
CKIP Tags	68.6	0.4	CKIP Tags	69.4	0.5
UPenn VA-Tag	75.1	0.4	UPenn VA-Tag	70.3	0.5
UPenn VV-Tag	74.4	0.4	UPenn VV-Tag	71.5	0.5
All Verb Tags	68.6	0.4	All Verb Tags	68.4	0.5
Peri. Part.	62.9	0.4	Peri. Part.	58.1	0.5
Pass. Part.	68.6	0.4	Pass. Part.	67.9	0.5
A-A Morph.	48.5	0.4	A-A Morph.	49.0	0.5
A-N Morph.	48.0	0.4	A-N Morph.	48.0	0.5
A-V Morph.	48.7	0.4	A-V Morph.	48.4	0.5
N-A Morph.	50.0	0.4	N-A Morph.	50.0	0.5
N-N Morph.	47.6	0.4	N-N Morph.	47.6	0.5
N-V Morph.	49.7	0.4	N-V Morph.	47.6	0.5
V-A Morph.	56.0	0.5	V-A Morph.	66.1	0.5
V-N Morph.	57.4	0.5	V-N Morph.	59.2	0.5
V-V Morph.	47.0	0.5	V-V Morph.	48.0	0.5
Sem. Spec.	60.5	0.5	Sem. Spec.	49.0	0.5
Avg. Length	47.8	0.5	Avg. Length	49.4	0.5

Table 5.2: Percent accuracy and standard error of the verb classification task using each Chinese feature individually, with 8-fold cross-validation training method repeated 50 times.

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Features	*74.7*	*0.7*	*All Features*	*74.2*	*0.8*
All – Causativity	74.7	0.7	All – Causativity	74.2	0.8
All – VBN	74.7	0.7	All – VBN	74.2	0.8
All – Passive Voice	74.9	0.7	All – Passive Voice	74.3	0.8
All – Transitivity	74.7	0.7	All – Transitivity	74.1	0.8
All – Animacy	74.9	0.7	All – Animacy	74.3	0.8

Table 5.3: Percent accuracy and standard error of the verb classification task by removing each individual English feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times.

from removing individual features. Table 5.3 compiles the results. The first line of this table shows the performance of classification using all English and Chinese features. All the features combined perform very well: all the English and aligned Chinese features combined have an accuracy of 74.7%, a reduction of 49.4% of the error rate; all the English and unaligned Chinese features combined have an accuracy of 74.2%, a reduction

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Features	*74.7*	*0.7*	*All Features*	*74.2*	*0.8*
All – CKIP Tags	76.6	0.7	All – CKIP Tags	64.0	0.8
All – UPenn VA-Tag	74.7	0.7	All – UPenn VA-Tag	74.1	0.8
All – UPenn VV-Tag	74.6	0.7	All – UPenn VV-Tag	74.2	0.8
All – All Verb Tags	58.6	0.7	All – All Verb Tags	56.1	0.8
All – Peri. Part.	72.8	0.7	All – Peri. Part.	74.4	0.8
All – Pass. Part.	62.0	0.7	All – Pass. Part.	65.9	0.8
All – A-A Morph.	74.7	0.7	All – A-A Morph.	74.5	0.8
All – A-N Morph.	74.9	0.7	All – A-N Morph.	74.6	0.8
All – A-V Morph.	74.7	0.7	All – A-V Morph.	74.2	0.8
All – N-A Morph.	74.7	0.7	All – N-A Morph.	74.2	0.8
All – N-N Morph.	74.7	0.7	All – N-N Morph.	74.2	0.8
All – N-V Morph.	74.7	0.7	All – N-V Morph.	74.2	0.8
All – V-A Morph.	74.7	0.7	All – V-A Morph.	74.2	0.8
All – V-N Morph.	75.2	0.7	All – V-N Morph.	74.5	0.8
All – V-V Morph.	74.7	0.7	All – V-V Morph.	74.2	0.8
All – Sem. Spec.	74.8	0.7	All – Sem. Spec.	77.8	0.8
All – Avg. Length	74.7	0.7	All – Avg. Length	74.7	0.8

Table 5.4: Percent accuracy and standard error of the verb classification task by removing each individual Chinese feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times.

of 48.4% of the error rate. On both panels, there is little to no difference between the performance using the full feature set and the performance of the same set with only one English feature removed. The results show that each of the English features contributes little to the verb classification.

Our next task was to perform 8-fold cross-validation on all features and all features with one Chinese feature removed. Results from this experiment are shown in Table 5.4. In both panels, we see a significant decrease in performance by removing either *All Verb Tags* or *Pass. Part.* (the shaded areas of the table). Using unaligned data, there is also a significant performance degradation from removing *CKIP Tags*. However, other Chinese features that have better-than-chance performance individually (see Table 5.2) do not affect the overall performance of the remaining features.

Note that the union of all *CKIP Tags* and all *UPenn Tags* forms *All Verb Tags*, which

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Eng. Features	41.3	0.7	All Eng. Features	Aligned = Unaligned	
All Chi. Features	75.4	0.7	All Chi. Features	74.1	0.7
All Features	74.7	0.7	All Features	74.2	0.8

Table 5.5: Percent accuracy and standard error of the verb classification task using all English-only features, all Chinese-only features, and the full bilingual feature set, with 8-fold cross-validation training method repeated 50 times.

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Eng. Features	*41.3*	*0.7*	*All Eng. Features*	*Aligned = Unaligned*	
+ CKIP Tags	77.5	0.7	+ CKIP Tags	77.9	0.8
+ UPenn VA-Tag	72.7	0.7	+ UPenn VA-Tag	55.4	0.8
+ UPenn VV-Tag	71.8	0.7	+ UPenn VV-Tag	54.8	0.8
+ All Verb Tags	67.3	0.7	+ All Verb Tags	63.3	0.8
+ Peri. Part.	57.6	0.7	+ Peri. Part.	50.9	0.8
+ Pass. Part.	66.5	0.7	+ Pass. Part.	60.6	0.8
+ A-A Morph.	41.1	0.7	+ A-A Morph.	43.8	0.8
+ A-N Morph.	45.0	0.7	+ A-N Morph.	41.3	0.8
+ A-V Morph.	41.9	0.7	+ A-V Morph.	44.1	0.8
+ N-A Morph.	41.3	0.7	+ N-A Morph.	44.9	0.8
+ N-N Morph.	40.8	0.7	+ N-N Morph.	41.7	0.8
+ N-V Morph.	44.8	0.7	+ N-V Morph.	45.8	0.8
+ V-A Morph.	55.7	0.7	+ V-A Morph.	52.7	0.8
+ V-N Morph.	45.9	0.7	+ V-N Morph.	45.9	0.8
+ V-V Morph.	32.2	0.7	+ V-V Morph.	43.1	0.8
+ Sem. Spec.	51.3	0.7	+ Sem. Spec.	45.1	0.8
+ Avg. Length	41.6	0.7	+ Avg. Length	40.6	0.8
All Chi. Features	75.4	0.7	All Chi. Features	74.1	0.7

Table 5.6: Percent accuracy and standard error of the verb classification task by augmenting the full English feature set with each individual Chinese feature, with 8-fold cross-validation training method repeated 50 times.

is a feature set of 15 + 2 = 17 POS tags. In the left panel, removing *All Verb Tags*, but neither *CKIP Tags* nor *UPenn Tags*, made an impact on the performance. In the right panel, removing *CKIP Tags* alone decreases the overall accuracy (significant at $p < .05$), but removing *All Verb Tags* increases the error rate even more. We interpret this to mean that both *CKIP Tags* and *UPenn Tags* divide the verb classes in a similar way, as

our linguistic analysis would lead us to expect. If one of the three features is missing, the remaining two can still provide similar information for verb classification.

Now let's look at the overall performance of English-only and Chinese-only features in Table 5.5. Combining only the English features has an accuracy of 41.3%, a below-chance performance. Combining only the Chinese features has an accuracy of 75.4% using the aligned method, and 74.7% using the unaligned method. They do not differ much from the accuracies achieved by using all features. This again shows that the English features are adding little information to the overall feature set.

Table 5.6 shows the performance of classification by augmenting individual Chinese features to all the English features. Note that many Chinese features improve the English performance. Those which improve performance are almost exactly the same set of features which have a higher-than-50% accuracy individually (see Table 5.2). Despite the fact that the English feature set has a less-than-chance accuracy, when combined with individual Chinese features, the performance exceeds those of using individual Chinese features. The most notable combination is the English features combined with *CKIP Tags*, where the accuracy shoots up to 78%, aligned or unaligned. Each Chinese feature individually has a maximum accuracy of only 75.1% (*UPenn VA-Tag*, aligned, see Table 5.2), and all the Chinese features combined have a maximum accuracy of 75.4% (aligned). Consider the combination *All English Features + CKIP Tags*. The improvement in performance over the feature sets using only monolingual features is statistically significant. No other combination of features, using strictly HKLaws data, exceeds 75%.[1]

[1]Other combinations tried and not presented here are: all Chinese features combined with individual English feature; various combinations of Chinese features; various combinations of English features.

Feature	%Accuracy	%SE
Causativity	47.8	0.4
VBN	49.0	0.4
Passive Voice	49.9	0.4
Transitivity	48.7	0.4
Animacy	72.5	0.4

Table 5.7: Percent accuracy and standard error of the verb classification task using each English feature individually, with 8-fold cross-validation training method repeated 50 times. English WSJ data used.

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Features	*65.3*	*0.6*	*All Features*	*71.5*	*0.6*
All – Causativity	73.2	0.8	All – Causativity	69.4	0.8
All – VBN	72.2	0.8	All – VBN	69.4	0.8
All – Passive Voice	72.7	0.8	All – Passive Voice	69.4	0.8
All – Transitivity	66.8	0.8	All – Transitivity	69.1	0.8
All – Animacy	74.9	0.8	All – Animacy	73.9	0.8

Table 5.8: Percent accuracy and standard error of the verb classification task by removing each individual English feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data.

5.2.2 WSJ Data

Now we turn to our experiments using WSJ data for the English features. With a few exceptions, our analysis revealed that English HKLaws data contributes little to our automatic learning task. However, WSJ data was considered useful for Merlo and Stevenson's (2001a) set of verbs. We decided to replicate their experiments using our English data (new verbs, and different extraction patterns), then we duplicated the bilingual experiments in Section 5.2.1 by pairing the WSJ data with Chinese HKLaws data.

Table 5.7 shows the performance of classification using individual English features. Similarly to the English HKLaws results in Table 5.1, most features perform no better than chance. However, using WSJ data, one feature, *Animacy*, does achieve an accuracy of 72.5% in distinguishing the unaccusative verbs from the object-drop verbs.

Table 5.8 shows the results using all subsets of English and Chinese features with one English feature removed using WSJ data. In the aligned panel, we see that, except for *Transitivity*, removing any of the features improves the overall performance. In the unaligned panel, none of the features make a statistically significant impact on the overall performance. It is unclear from this experiment whether the English features extracted from the WSJ contribute much information to the Chinese features, despite the fact that one feature, *Animacy*, has above-chance performance individually.

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Features	*65.3*	*0.6*	*All Features*	*71.5*	*0.6*
All – CKIP Tags	72.5	0.7	All – CKIP Tags	65.0	0.8
All – UPenn VA-Tag	64.7	0.7	All – UPenn VA-Tag	69.0	0.8
All – UPenn VV-Tag	64.8	0.7	All – UPenn VV-Tag	69.0	0.8
All – All Verb Tags	60.8	0.7	All – All Verb Tags	65.0	0.8
All – Peri. Part.	72.3	0.7	All – Peri. Part.	70.8	0.8
All – Pass. Part.	70.0	0.7	All – Pass. Part.	65.2	0.8
All – A-A Morph.	72.1	0.7	All – A-A Morph.	69.2	0.8
All – A-N Morph.	72.3	0.7	All – A-N Morph.	69.1	0.8
All – A-V Morph.	72.9	0.7	All – A-V Morph.	69.1	0.8
All – N-A Morph.	72.3	0.7	All – N-A Morph.	69.1	0.8
All – N-N Morph.	72.2	0.7	All – N-N Morph.	69.0	0.8
All – N-V Morph.	72.3	0.7	All – N-V Morph.	69.1	0.8
All – V-A Morph.	72.4	0.7	All – V-A Morph.	69.8	0.8
All – V-N Morph.	72.8	0.7	All – V-N Morph.	69.2	0.8
All – V-V Morph.	74.0	0.7	All – V-V Morph.	71.6	0.8
All – Sem. Spec.	73.1	0.7	All – Sem. Spec.	69.4	0.8
All – Avg. Length	72.6	0.7	All – Avg. Length	69.2	0.8

Table 5.9: Percent accuracy and standard error of the verb classification task by removing each individual Chinese feature from a full bilingual feature set, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data.

Table 5.9 presents results from similar experiments for Chinese features. In the left panel, with the exception of *All Verb Tags*, none of the Chinese features seems to decrease the overall accuracy (line 1). However, in the right panel, removing each of *CKIP Tags*, *All Verb Tags*, and *Pass. Part.* yields a decrease of 5–7% in performance. These are the

same three features which made a difference using solely HKLaws data (see Table 5.4).

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Eng. Feat.	*66.3*	*0.6*	*All Eng. Feat.*	*Aligned = Unaligned*	
+ CKIP Tags	72.1	0.6	+ CKIP Tags	75.6	0.6
+ UPenn VA-Tag	80.6	0.6	+ UPenn VA-Tag	75.0	0.6
+ UPenn VV-Tag	78.6	0.6	+ UPenn VV-Tag	73.2	0.6
+ All Verb Tags	77.1	0.6	+ All Verb Tags	69.5	0.6
+ Peri. Part.	66.2	0.6	+ Peri. Part.	76.2	0.6
+ Pass. Part.	64.9	0.6	+ Pass. Part.	70.1	0.6
+ A-A Morph.	66.1	0.6	+ A-A Morph.	65.4	0.6
+ A-N Morph.	66.3	0.6	+ A-N Morph.	66.3	0.6
+ A-V Morph.	67.1	0.6	+ A-V Morph.	66.9	0.6
+ N-A Morph.	66.3	0.6	+ N-A Morph.	66.3	0.6
+ N-N Morph.	66.8	0.6	+ N-N Morph.	66.8	0.6
+ N-V Morph.	65.8	0.6	+ N-V Morph.	63.8	0.6
+ V-A Morph.	66.8	0.6	+ V-A Morph.	65.6	0.6
+ V-N Morph.	65.3	0.6	+ V-N Morph.	65.3	0.6
+ V-V Morph.	64.4	0.6	+ V-V Morph.	64.5	0.6
+ Sem. Spec.	62.7	0.6	+ Sem. Spec.	63.6	0.6
+ Avg. Length	66.1	0.6	+ Avg. Length	66.1	0.6
All Chi. Feat.	75.4	0.6	All Chi. Feat.	74.1	0.6

Table 5.10: Percent accuracy and standard error of the verb classification task by augmenting all the English features with each individual Chinese feature, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data.

We next look at the performance of individual Chinese features by adding them to the set of English features. In the left panel of Table 5.10, the addition of various combinations of verb POS tags improves the accuracy of the English features by at least 5.8%. In the right panel, the addition of these features and the external particles also improves performance. Note that the best English and Chinese feature combinations (i.e., Aligned: *All Eng. Feat. + UPenn VA-Tag*; Unaligned: *All Eng. Feat. + Peri. Part.*) also perform statistically significantly better than the Chinese-only features (Aligned: 75.4%; Unaligned: 75.1%).

Table 5.11 compares the performance between English-only and Chinese-only features, and all features. The performance of the Chinese-only features exceeds the performance

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Eng. Features	66.3	0.6	All Eng. Features	Aligned = Unaligned	
All Chi. Features	75.4	0.7	All Chi. Features	74.1	0.7
All Features	65.3	0.6	All Features	71.5	0.6

Table 5.11: Percent accuracy and standard error of the verb classification task using all English-only features, all Chinese-only features, and all features, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data.

Aligned Features	%Accuracy	%SE	Unaligned Features	%Accuracy	%SE
All Chi. Feat.	*75.4*	*0.6*	*All Chi. Feat.*	*74.1*	*0.6*
+ Causativity	75.3	0.6	+ Causativity	74.1	0.8
+ VBN	75.4	0.6	+ VBN	73.9	0.8
+ Passive Voice	75.1	0.6	+ Passive Voice	74.1	0.8
+ Transitivity	75.2	0.6	+ Transitivity	74.2	0.8
+ Animacy	67.8	0.6	+ Animacy	71.0	0.8

Table 5.12: Percent accuracy and standard error of the verb classification task by augmenting all the English features with each individual Chinese feature, with 8-fold cross-validation training method repeated 50 times. English WSJ data augmented by Chinese HKLaws data.

of the other two sets of features. This shows that simply combining all the English and Chinese features does not always improve performance. To find out which particular English features decrease the performance of the Chinese-only features, we did an experiment adding individual English feature to the Chinese features.

Table 5.12 shows that *Animacy* is the culprit. Observe that adding this feature yields a decrease in performance of 3–8%. This shows that this feature does not work well with the combination of all Chinese features. However, we also see that the performance of various combinations of English and Chinese features (sometimes including *Animacy*) is comparable to the performance of using Chinese features alone. Further experiments finding the best combination(s) are needed.

5.2.3 Summary of Cross-Validation Results

Now we have presented cross-validation results using HKLaws data and WSJ data. Using these two sets of data, we examined the following feature combinations:

- individual English features
- individual Chinese features
- removing individual English features from a full bilingual feature set
- removing individual Chinese features from a full bilingual feature set
- adding individual Chinese features to a full English feature set

Evaluating Individual English Features As our manual data analysis (Section 4.2.2) would lead us to expect, English features are not very useful in our classification task. Using HKLaws data, none of the individual English features has an above-chance performance (see Table 5.1). Similarly, we observe little to no performance degradation by removing individual English features (see Table 5.3). Using WSJ data, the analysis is not as trivial. Except *Animacy*, none of the individual English features has an above-chance accuracy (see Table 5.7). Similarly, removing individual English features from the full bilingual feature set shows no performance degradation (see Table 5.8). In fact, the removal of some English features, including *Animacy*, shows a performance gain. Using WSJ data, it is unclear how to assess the usefulness of the English feature, *Animacy*. In general, contrary to Merlo and Stevenson's (2001a) findings, we do not find English features alone useful in our classification task.

Evaluating Individual Chinese Features Unlike the results using English-only features, Chinese features perform very well. Many individual Chinese features have an above-chance performance (see Table 5.2). However, by removing individual Chinese

features from the full set of bilingual features, fewer of them show a performance degradation; specifically, only the removal of Chinese POS tags and external particles decreases accuracy (see Tables 5.4 and 5.9). In sum, Chinese features are more useful in distinguishing unaccusative verbs from object-drop verbs.

Evaluating Multilingual Features Recall that our ultimate goal is to observe how multilingual data influence the automatic learning of verb classification. We do so by assessing the potential performance gain by adding individual Chinese features to a set of English features. The addition of many of the individual Chinese features indeed increases the performance of English features (see Tables 5.6 and 5.10). Furthermore, the best multilingual accuracy exceeds the best monolingual accuracy. Interestingly, while some (combinations of) features do not exceed chance performance (e.g., all English-only features using HKLaws data), in the "right" combination, they can still contribute to the learning of verb classification. For example, using HKLaws data, the performance of the combination of English features and CKIP Tags exceeds the performance of using CKIP Tags only. In conclusion, multilingual features provide performance benefits over monolingual features.

5.3 Results Using Leave-One-Out Methodology

In this section, we report results from our experiments using the leave-one-out method. Recall that in each iteration, we leave one vector out for testing and use the remaining vectors for training. Although this approach is similar to the N-fold cross-validation methodology with $N = 32$, we do not consider it redundant. This approach is useful in finding the best set of features in classifying a particular class because we know the classification result of each individual verb and hence the precision and recall by verb class. The performance on individual verbs as well as precision and recall per class provide alternative ways of evaluation other than average accuracy.

From the previous section, we know that using all Chinese-only features yields an accuracy of above 70%; we also know that some combination of English and Chinese features has a comparable performance. (In this section, we only report results using English WSJ data in conjunction with the Chinese HKLaws data. We found that our English data collection method does not work well with the English HKLaws corpus. See section 4.2 for a discussion.) As we mentioned earlier in this chapter, we have a total of 21 groups of features (35 features if we consider each POS tag as one feature). Performing an exhaustive search for the best combinations is difficult. Knowing that some features evidently contribute to learning, we select for our leave-one-out experiments only those features that were useful in our cross-validation experiments. These features are:

- All Chinese Features
- All English Features + a combination of CKIP Tags, Passive Particles, Periphrastic Particles
- All English Features – Animacy + a combination of CKIP Tags, Passive Particles, Periphrastic Particles

For each feature set we used, we were able to calculate the precision and recall for each class. Since our task is an exhaustive binary classification task, the recall and precision of one class directly affect the precision and recall of the other class (e.g., a high number of false negatives in class A results in a low recall for class A and a low precision in class B). Hence we also calculated a balanced F-score (as $2PR/(P+R)$, where P and R are precision and recall) for each class and the percent accuracy over all verbs.

In Table 5.13, using F-score as a performance measure, we see that the Chinese features perform almost equally well on either verb class. Using the aligned data, the Chinese features perform slightly better on object-drop verbs than on unaccusative verbs (by producing fewer false negatives in classifying object-drop verbs). Using the unaligned

Features	Unaccusative			Object-Drop			All Verbs
	Recall	Precision	F-score	Recall	Precision	F-score	%Accuracy
Aligned	0.75	0.80	0.77	0.81	0.76	0.79	78.1
Unaligned	0.88	0.78	0.82	0.75	0.86	0.80	81.2

Table 5.13: Recall, precision, balanced F-score, and percent accuracy of the verb classification task using all Chinese features, with leave-one-out training method.

data, the Chinese features perform slightly better on unaccusative verbs than on object-drop verbs (by producing fewer false negatives in classifying unaccusative verbs).

Now we turn to the results using a selection of English and Chinese features. We highlight the features with the best overall performance in Table 5.14. We find that many combinations that include *CKIP Tags* perform very well. In comparison to the performance using all English features, all performance measures indicate that these combinations have a better performance. In comparison to the performance using all Chinese features, the analysis is slightly more complicated. The recall on unaccusative verbs either remains the same or is slightly worse but there is also a jump in the precision (slightly more false negatives but fewer false positives). Similarly, the precision on object-drop verbs is either the same or slightly worse but there is also a jump in the recall (slightly more false positives but fewer false negatives). There is reduction in the total number of errors resulting in a better overall accuracy. On a closer examination of the table, the addition of *CKIP Tags* alone gives the best F-score on unaccusative verbs in the aligned (top) panel. However, to get the best F-score on object-drop verbs, the addition of external particles is needed. In the unaligned panel, the addition of *CKIP Tags* and *Passive Particles* has the best overall performance. This feature set has the best recall, precision (and therefore F-score) on either class, and the best overall accuracy. In general, multilingual features perform better than monolingual features.

From section 5.2.2, we know that the feature *Animacy* does not always work well with other features. Hence we duplicated the above experiment with *Animacy* removed

Features	Unaccusative			Object-Drop			All Verbs
	Recall	Precision	F-score	Recall	Precision	F-score	%Accuracy
All Eng.	0.63	0.63	0.63	0.63	0.63	0.63	62.5
Aligned							
+ 1	0.75	0.86	0.80	0.88	0.78	0.82	81.3
+ 2	0.56	0.60	0.58	0.63	0.59	0.61	59.4
+ 3	0.50	0.53	0.52	0.56	0.53	0.55	53.1
+ 1,2	0.69	0.92	0.79	0.94	0.75	0.83	81.3
+ 2,3	0.44	0.54	0.48	0.63	0.53	0.57	53.1
+ 1,3	0.69	0.92	0.79	0.94	0.75	0.83	81.3
+ 1,2,3	0.69	0.92	0.79	0.94	0.75	0.83	81.3
Unaligned							
+ 1	0.63	0.63	0.63	0.63	0.63	0.63	62.5
+ 2	0.69	0.79	0.73	0.81	0.72	0.76	75.0
+ 3	0.75	0.86	0.80	0.88	0.78	0.82	81.3
+ 1,2	0.75	0.92	0.83	0.94	0.79	0.86	84.4
+ 2,3	0.69	0.69	0.69	0.69	0.69	0.69	68.8
+ 1,3	0.50	0.67	0.57	0.75	0.60	0.67	62.5
+ 1,2,3	0.50	0.80	0.62	0.88	0.64	0.74	68.8

Table 5.14: Recall, precision, balanced F-score, and percent accuracy of the verb classification task using all English features and a combination of Chinese features, with leave-one-out training method. (1 = CKIP Tags; 2 = Passive Particles; 3 = Periphrastic Particles)

from all the feature sets. The results are shown in Table 5.15. The set including *CKIP Tags* and *Passive Particles* has the best overall accuracy on the aligned and unaligned data. By removing *Animacy*, only some performance indicators show that this feature set has a slightly better performance than the best feature sets with *Animacy*. The tradeoff between precision and recall is more balanced with *Animacy* removed. Hence, the differences between the two sets of F-scores are not huge. In general, with or without *Animacy*, adding Chinese features has a performance benefit over monolingual features. Clearly, we did not exhaustively find the best possible feature set(s), but we have shown that the performance of a multilingual feature set is superior.

Since we have accuracy information for each verb, we also want to find out if some verbs are consistently incorrectly classified. We compared the set of misclassified verbs

Features	Unaccusative			Object-Drop			All Verbs
	Recall	Precision	F-score	Recall	Precision	F-score	%Accuracy
All Eng.–Anim.	0.63	0.63	0.63	0.63	0.63	0.63	62.5
Aligned							
+ 1	0.81	0.72	0.76	0.69	0.79	0.73	75.0
+ 2	0.69	0.73	0.71	0.75	0.71	0.73	71.9
+ 3	0.31	0.83	0.45	0.94	0.58	0.71	62.5
+ 1,2	0.81	0.87	0.84	0.88	0.82	0.85	84.4
+ 2,3	0.69	0.69	0.69	0.69	0.69	0.69	68.8
+ 1,3	0.75	0.75	0.75	0.75	0.75	0.75	75.0
+ 1,2,3	0.75	0.75	0.75	0.75	0.75	0.75	75.0
Unaligned							
+ 1	0.88	0.74	0.80	0.69	0.85	0.76	78.1
+ 2	0.50	0.73	0.59	0.81	0.62	0.70	65.6
+ 3	0.75	0.57	0.65	0.44	0.64	0.52	59.4
+ 1,2	0.88	0.82	0.85	0.81	0.87	0.84	84.4
+ 2,3	0.50	0.73	0.59	0.81	0.62	0.70	65.6
+ 1,3	0.81	0.72	0.76	0.69	0.79	0.73	75.0
+ 1,2,3	0.81	0.72	0.76	0.69	0.79	0.73	75.0

Table 5.15: Recall, precision, balanced F-score, and percent accuracy of the verb classification task using all English features - Animacy + a combination of Chinese features, with leave-one-out training method. (1 = CKIP Tags; 2 = Passive Particles; 3 = Periphrastic Particles)

between feature sets. No two sets of features share exactly the same set of incorrectly classified verbs. Contrary to Merlo and Stevenson's (2001a) analysis, we found almost all verbs were misclassified in one or more feature sets. Only two unaccusative verbs (*clear* and *decrease*) and two object-drop verbs (*paint* and *compose*) were always correctly classified. None of them has the highest or lowest frequency in the WSJ or the HKLaws corpus. Out of all 32 verbs, one verb, *produce*, was incorrectly classified in all experiments, though *produced* has the highest Chinese frequency (but it does not have the highest English frequency). Therefore, we cannot conclude whether the frequency of a verb is correlated with the accuracy of its classification.

Experiment	Data	Features	%Acc.	%SE	Table
English Only	English HKLaws	All Eng. Features	41.3	0.7	5.5
		Transitivity	49.5	0.5	5.1
	WSJ	All Eng. Features	66.3	0.6	5.11
		Animacy	72.5	0.4	5.7
Chinese Only	Aligned HKLaws	All Chi. Features	75.4	0.7	5.5
		UPenn VA-Tag	75.1	0.4	5.2
	Unaligned HKLaws	All Chi. Features	74.1	0.7	5.5
		UPenn VV-Tag	71.5	0.5	5.2
Chinese and English	Aligned HKLaws + Eng. HKLaws	All Features	74.7	0.7	5.5
		All Eng. Feat. + CKIP Tags	77.5	0.7	5.6
	Unaligned HKLaws + Eng. HKLaws	All Features	74.2	0.7	5.5
		All Eng. Feat. + CKIP Tags	77.9	0.8	5.6
	Aligned HKLaws + WSJ	All Features	65.3	0.6	5.11
		All Eng. Feat. + UPenn VA-Tag	80.6	0.6	5.10
	Unaligned HKLaws + WSJ	All Features	71.5	0.6	5.11
		All Eng. Feat. + Peri. Part.	76.2	0.6	5.10

Table 5.16: Summary of the best feature combinations using N-fold cross-validation training methodology

5.4 Summary of Results

In this chapter, we documented results using two types of training methods: N-fold cross-validation and leave-one-out methodologies. In the N-fold cross-validation experiments, we tested different combinations of features from the HKLaws corpus and the WSJ: English features alone, Chinese features alone, and a combination of English and Chinese features. Recall that we selected 16 unaccusative verbs and 16 object-drop verbs for our task, hence the chance performance is 50%.

Using the HKLaws corpus only, contrary to Merlo and Stevenson's findings, we did not find English features useful. That is, using an English corpus 10% of the size of the corpus used by Merlo and Stevenson (2001a), the best accuracy using English features alone is no better than chance performance (49.5% accuracy, SE 0.5%). Chinese features alone

For Unaccusative Verbs Only					
Corpus Data	Features	Rec.	Prec.	F	Table
Aligned HKLaws + WSJ	Eng. – Anim. + 1 + 2	0.81	0.87	0.84	5.15
Unaligned HKLaws + WSJ	Eng. + 1 + 2	0.75	0.92	0.83	5.14,
	Eng. – Anim. + 1 + 2	0.88	0.82	0.85	5.15
For Object-Drop Verbs Only					
Corpus Data	Features	Rec.	Prec.	F	Table
Aligned HKLaws + WSJ	Eng. + 1	0.94	0.75	0.83	5.14,
	Eng. + 1 + 2				5.15
	Eng. + 1 + 3				
	Eng. + 1 + 2 + 3				
	Eng. – Anim. + 1 + 2	0.88	0.82	0.85	
Unaligned HKLaws + WSJ	Eng. + 1 + 2	0.94	0.79	0.86	5.14,
	Eng. – Anim. + 1 + 2	0.81	0.87	0.84	5.15
All Verbs					
Corpus Data	Features	%Accuracy			Table
Aligned HKLaws + WSJ	Eng. – Anim. + 1 + 3	84.4			5.15
Unaligned HKLaws + WSJ	Eng. – Anim. + 1 + 3	84.4			5.15

Table 5.17: Summary of the best feature combinations using leave-one-out training methodology (1 = CKIP Tags; 2 = Passive Particles; 3 = Periphrastic Particles)

performed better than the English features. Using aligned Chinese data only, the best accuracy is 75.1% with SE 0.4%; using unaligned Chinese data only, the best accuracy is 71.5% with SE 0.4%. That is, Chinese features alone achieved an error reduction rate of at most 50.2% (of the baseline error rate of 50%). The performance achieved by combining English and Chinese features is better than the Chinese features (77.5% with SE 0.7% using aligned data and 77.9% with SE 0.8% using unaligned data). In general, although the unaligned feature collection method introduces more noise to the data, using the unaligned dataset, we do not find the accuracy consistently worse.

Using WSJ data for our 32 verbs, we only found one useful English feature, *Animacy*, achieving an accuracy of 72.5% with SE 0.4%. Despite that, combining all English features has an above-chance performance (66.3% accuracy, SE 0.6%). Combining the English features extracted from the WSJ and the Chinese features from the HKLaws

corpus, the best performance is at 80.6% with SE 0.6%. This shows a significant improvement over monolingual (and even multilingual) features from either corpus alone.

On the basis of the results from the N-fold cross-validation experiments, we conducted leave-one-out tests using Chinese data from the HKLaws corpus and English data from the WSJ. In the leave-one-out experiments, we varied the precise set of Chinese features used. In conjunction with the English features, the best feature sets outperformed the monolingual features as well. This again shows there is a considerable improvement appending the Chinese features to the existing English features.

Tables 5.16 and 5.17 give a summary of the feature combinations with the best performance in our experiments. In the next chapter, we will discuss the implication of pairing cross-linguistic data.

Chapter 6

Discussion

In Chapter 5, we made use of parallel and monolingual corpora to see that Chinese features, alone or in combination, can be useful in automatic classification of English verbs. We have also seen that when they are combined with English features, the performance is better than the performance of using English features alone. Based on these preliminary results, we have the following observations:

1. There are some Chinese verb features showing syntactic and semantic distinctions between semantic classes of English verbs.

2. These Chinese features are surface syntactic features which make them easy to detect in a corpus, while in English, similar information is implicit in the meaning.

3. In a parallel English-Chinese corpus, differences in the English verb usages across the classes give rise to distributional differences of these Chinese syntactic features in the translation.

4. The distributional differences of the Chinese features are useful in classifying English verbs.

5. Using non-parallel bilingual corpora, these distributional differences can be useful in automatic classification of English verbs as well.

We will discuss these observations in the subsequent sections. Specifically we will look at the contributions of Chinese verb features and the use of multilingual corpora in automatic verb classification.

6.1 Chinese Lexical/Sublexical Features and Verb Classification

Although we could not find any unergative (manner-of-motion) verbs in our HKLaws corpus, in section 3.3, we made predictions that our proposed statistical features should at least provide a two-way distinction between unaccusative and object-drop verbs. We reiterate these predictions in Table 6.1.

Chinese Feature	Expected Frequency Pattern
(Stative) POS tag	Unerg, ObjDrop < Unacc
Periphrastic particles	Unerg, ObjDrop < Unacc
Passive particles	Unerg < Unacc < ObjDrop
Morpheme patterns: resultative constructions (V-A and V-V)	ObjDrop < Unerg < Unacc
Morpheme patterns: semantic specificity	ObjDrop < Unerg < Unacc
Morpheme patterns: average morpheme length	ObjDrop < Unerg < Unacc

Table 6.1: The Chinese features and their expected behaviour.

6.1.1 Individual Feature Performance

Our results confirm that many of these Chinese features, when used alone, can provide an above-chance, two-way distinction between unaccusative and object-drop verbs. However, there are two unexpected results. First, we found that the V-V combination is not a useful feature, while the V-N combination turns out to have a higher-than-50% accuracy when used individually. Though we decided to collect data on all morpheme

combinations, including the V-N combination, we find conflicting linguistic evidence as to whether this type of compound construction reflects the differences between verb classes. For example, Her (1996, 1997) found that many V-O compounds (equivalent to the V-N notation we use here) are non-compositional. Hence the meaning of a compound, and, as a result, its semantic classification cannot be derived based on the individual components. On the other hand, in a study on Chinese verbs (of emotion), Chang et al. (1999) suggested that V-N compounds, instead of V-V or V-A compounds, are preferred when expressing change-of-state events:

> In VV, the concept of an event is "diffused" after combining two similar events, since speaker[s] will extract the common attributes of the pair. It is common morpho-lexical strategy in Mandarin to concatenate two antonyms or synonyms to form the concept of "kind" or "property" ... [Hence, it is] natural for the VV compounds to be chosen to indicate a homogeneous state, but awkward to indicate an inchoative state.

Given the different linguistic views, the interaction between the sublexical components of a verb could be more complex than we have anticipated. Further research on Chinese verb compounds is needed.

Another unexpected result is that, although passive particles were useful in distinguishing the unaccusative verbs from the object-drop verbs, we found that passive particles co-occur with unaccusative verbs more often than with object-drop verbs, contrary to the predicted frequency patterns (see Table 6.1). This finding contradicts the hypothesis that unaccusative verbs occur less often in English passive voice sentences, and therefore result in fewer Chinese passive voice constructions. We postulate that the co-occurrence frequency of Chinese passive particles is related to the adjectival nature of unaccusative (change-of-state) verbs. The adjectival use is a type of passive use. For instance, "The *closed* door" implies "The door is *closed*." It is possible that there is a correlation between

verb adjectivization and the use of Chinese passive particles, but our original hypothesis is not sufficient to account for this phenomenon.

6.1.2 Contribution of Chinese Features to English Verb Classification

In combination with English features, we have found that Chinese POS tags, passive particles, and occasionally periphrastic particles, work better than other Chinese features in improving the performance of English-only features. Our explanation is twofold:

Passive and Periphrastic Particles We postulated that the behaviour of passive particles and periphrastic particles is correlated with the English passive voice and causativity features respectively: the use of passive particles is related to the English passive construction and the use of periphrastic particles is related to the causative alternation of English unergative and unaccusative verbs. In other words, these two Chinese features are not orthogonal to the English features. Despite that, given our English data extracted from the English HKLaws corpus and the WSJ, neither English feature proved to be useful. We believe the two Chinese features compensate for what is missing from the English features. (The Chinese features work as "backup" features.)

The notion of "backup" information also appears in (Merlo and Stevenson, 2001a). In their work, the English syntactic features, VBN POS tag and passive voice, are related to the feature, transitivity. These are not orthogonal features and yet by removing one of these features, there is a performance degradation. (At least this is true for the VBN POS tag and transitivity.) Although these features are related, none of these features capture exactly the same information. As the authors noted, " ...the counts will be imperfect approximations to the thematic knowledge, beyond the inevitable errors due to automatic extraction from large automatically annotated corpora." By including slightly different but related features, one feature may contribute information that may be imperfectly

extracted in the other two features (Suzanne Stevenson, personal communication). Here we believe the backup effect of passive and periphrastic particles is similar.

Chinese POS-Tags Unlike the previous two features, Chinese POS tag is a feature that does not overlap with any existing English features. Both the CKIP and the UPenn annotation guidelines broadly classify Chinese verbs as either state or action verbs. As discussed in Chapter 3, in Chinese, the state vs. action distinction is related to whether a verb can be adjectivized. This property is not in conflict with a related syntactico-semantic property in English verbs. We saw that optionally transitive unaccusative verbs are change-of-state verbs in which the final state is implicit in the meaning of the verb. In English, the change-of-state meaning is sometimes manifested syntactically in the passive adjectival form. In this case, the passive adjectival form has a stative reading. For example, in "a frozen river" the water in the river must be solid; in "a burnt toast" the toast must be black (Verspoor, 1997). Although Chinese is less restricted than English on which classes of verbs can be adjectivized, here it seems to be useful in distinguishing unaccusative verbs from the object-drop verbs.

Although we found the Chinese POS tags useful in our task, one concern we have is the reliability of the CKIP POS tagger we used and the "correctness" of the CKIP POS annotation guideline. Thus far we found exactly one paper documenting the accuracy of this POS tagger (Liu et al., 1995). The documented accuracy is between 96% and 98% (on a two-million character corpus). This is a comparable accuracy to some of the existing English POS taggers (e.g., Brill, 1993; Ratnaparkhi, 1996). Despite this, we are aware that it is not clear if there is a direct relationship between the accuracy of the CKIP POS tagger and the usefulness of the POS tags, which leads us to our second point. Tsao (1996) heavily criticized the state-action dichotomy in the CKIP verb classification as "strange" by assigning some non-state verbs, such as the Chinese equivalent of *borrow*, as state verbs. Although the state-action distinction seems to divide the unaccusative

and the object-drop verbs nicely at this point, the usefulness of the CKIP classification is questionable. Further research on the state-action/adjectival nature of (unaccusative) verbs is needed.

Note that we are not claiming that the Chinese features chosen for this study provide a relevant level of representation of Chinese verbs. Despite some claims that unergative verbs are distinguishable syntactically from unaccusative verbs in Chinese (Yuan, 1999), we are not claiming that Chinese verbs can be clustered the same way as English verbs in Levin's standard (1993) either. Instead, we want to show that a particular non-English feature is useful if it provides a multi-way distinction between semantic verb classes for English. The strategy of learning about some property in one language using a property in another language is not unlike the "substitution" strategy in SLA that we mentioned in Chapter 2. As suggested by Helms-Park (1997), L1 transfer occurs "upon perceiving an overlap between L1 and L2". Obviously, it is unlikely there is a complete overlap of features between two languages, hence not all L1 transfer effects in humans are positive. For example, in the studies we cited in Section 2.1, depending on the context of the experiments, only some L1-to-L2 transfer effects are positive (e.g., Inagaki, 1997; Wang and Lee, 1999). Similarly, only some non-English features are useful in our machine-learning setting.

The construction of our experiments is inspired by SLA research, and we found that many Chinese features are useful in distinguishing our English verb classes. However, that is not to say our Chinese features are exactly the features that help ESL learners in acquiring English verb classes.[1] Instead, for the purpose of our classification task, we believe that a feature that is responsible for some positive transfer effects in humans tends to be a good candidate feature. For example, Mandarin Chinese has fewer restrictions

[1]As implied in Chapter 3, the Chinese features were not selected based on any evidence in SLA research. For example, the author found little to no research suggesting any transfer effects of the use of external particles and sublexical components, and yet some of our results show that they can be useful in our machine-learning task.

on what type of verbs can be adjectivized. Despite that, in some experiments, even low proficiency ESL learners were able to distinguish change-of-state verbs from the other verbs (Wang and Lee, 1999). Our earlier criticism of the CKIP POS tagset aside, our results confirm that stative/action verb tags in the Chinese POS tagset is a useful feature.

6.2 The Use of Multilingual Corpora

Merlo and Stevenson (2001a) used monolingual corpora. In our experiments, we used two corpora: The HKLaws parallel corpus and the monolingual WSJ corpus. As seen in Chapter 5, we paired up the data in the following ways:

- English HKLaws with Chinese HKLaws, sentence alignment followed.
- English HKLaws with Chinese HKLaws, no alignment used.
- The WSJ with Chinese HKLaws, aligned Chinese data used.
- The WSJ with Chinese HKLaws, unaligned Chinese data used.

The bitext-based technique is certainly not new. For example, Fung (1998) and Melamed and Marcus (1998) used a bilingual corpus to extract bilingual lexical entries. The assumption is that the bilingual corpus is sentence or segment alignable. The advantage of using an aligned/alignable corpus is that we can calculate some co-occurrence score between any two possible translations: one common theme in the work of these researchers is that given any arbitrary pair of tokens and some text coordinate system, the closer the two tokens' coordinates are, the more likely they are translational equivalents. The implication is that in one subcorpus of some bitext, the distribution of the different senses and usages of a word should be reflected or correlated in the distribution of its translations in the other subcorpus. We observe that some English syntactic/semantic constructions affect how a sentence is translated. We interpret this observation as (indirect) empirical evidence of the above hypothesis. For our work, we have suggested on

numerous occasions that some Chinese features are related to some English feature(s). The correlation between the distributions of a pair of translations (in a bitext) indirectly entails that for any pair of related features, one Chinese and one English, the Chinese feature can be used as a "supplementary" feature or even an approximation of the correlated English feature, and vice versa.[2]

Although there are more and more parallel corpora available for research purposes, most of them are considerably smaller in size than the more popular monolingual corpora (e.g., the WSJ). For example, our HKLaws corpus is only about one-tenth the size of the corpus used by Merlo and Stevenson (2001a). Although we do not know what a reasonable corpus size is for sufficient data, it is still possible that we have a data sparseness problem. One option is to use multiple non-parallel corpora. Thus far, we have found surprisingly few studies justifying the use of non-parallel texts in automatic learning (Fung and McKeown, 1997; Fung, 1998; Fung and Lo, 1998). From an SLA point of view, one justification is that L2 learners do not expose themselves to parallel text when acquiring a new language. Instead, they are usually exposed to one language at any point in time. From a methodological point of view, unaligned data is a possible substitute for aligned data. Although the unaligned data is certainly "less clean" than the aligned data since the distribution of the word senses and usages are no longer preserved, our results show that unaligned data has performance comparable to, if not better than, aligned data. Our work suggests that using multiple non-parallel monolingual corpora provides an alternative when large parallel corpora cannot be found.

[2]Our results show that the English passive voice feature is not a useful feature, but the Chinese passive particle feature is. Merlo and Stevenson (2001a) also did not find the English passive voice feature useful when it was combined with other features. However, this is not to say the English passive voice feature is not useful in general. (It is likely that our noisy extraction technique contributes to its poor performance.) On the contrary, the usefulness of Chinese passive particles may be indirect evidence that the English passive voice feature is useful. The argument is similar for the English causativity feature and Chinese periphrastic particles. Clearly, in machine-learning, the suggestion that a "supplementary" feature in one language can replace other features in another language is purely speculative. Further research, linguistic or otherwise, is necessary.

Chapter 7

Conclusions

We set out to investigate the notion of "L1 transfer" – the influence of knowledge of one language on the learning of another – being carried over to the machine learning setting. We have succeeded in showing that statistics of (carefully selected) multilingual features, collected from a bilingual corpus, are useful in automatic lexical acquisition in English – that is, they contribute positively to learning in our experiments. Our work is not the first study to utilize multilingual resources. However, we are one of the first to try a bilingual corpus-based technique for automatic lexical acquisition, and the first to address the particular problem of verb classification. Not only have we shown the usefulness of multilingual features, but like many studies before us (see Section 2.2), we have produced evidence that statistical distributions of various syntactic features capture semantic information about classes of verbs. Therefore we confirm that there is a connection between the syntax of verbs and their meaning.

In this thesis, we have presented results showing that a verb classification task using a multilingual corpus-based technique can achieve performance comparable to, and sometimes better than, using a monolingual corpus alone. We see our method as a first step in applying second language acquisition (SLA) phenomena in automatic lexical acquisition, especially from the point of view of an English-as-a-Second-Language (ESL)

student. For our work, we applied the "L1 transfer" phenomenon in the specific case of the first language (L1) being Chinese and the second language (L2) being English.

Despite our positive results, our work is very preliminary. As alluded to in earlier chapters, there are many areas that deserve further investigation. For instance, there are many aspects of SLA research we have not touched on which may contribute to the refinement of our work. We also have not considered using more-automatic techniques in extracting bilingual translation lexicons. For example, in Chapters 3 and 4, some part of the mining of the translations of our English target verbs was done by hand. In the following sections, we discuss our contributions and limitations in more detail.

7.1 Using Multilingual Corpora for Automatic Learning in English

7.1.1 Contributions

Merlo and Stevenson (2001a) used the ACL/DCI corpus. This corpus, which includes the Brown Corpus and years 1987-1989 of the WSJ, has a total size of 65 million words. To extract sufficiently discriminating statistics for the verb classification task, a relatively large monolingual corpus is needed. Other researchers have since replicated the same experiment using a smaller corpus (23 million words) from the WSJ, achieving similar accuracy levels (Anoop Sakar, private communication). We were not successful in replicating the results using an even smaller English corpus of 6.5 million words (the English subcorpus of HKLaws). However, in conjunction with a 9 million character Chinese corpus (the Chinese subcorpus of HKLaws), the combined data set contains enough information to produce a good performance, with an error reduction rate of as much as 56% (or an accuracy of 78%; see Table 5.6 in Section 5.2.1). Note that even the combined size of the two HKLaws sub-corpora is only two-thirds the size of the smallest corpus

attempted in previous automatic learning experiments for this problem. Although we do not know the lower bound of the corpus size necessary to provide sufficient data, we have pushed the limit lower than ever. This thesis has succeeded in showing that a smaller parallel corpus is a good substitute for, or a good addition to, a larger monolingual corpus.

7.1.2 Future Work: Corpus Size and Genre

In comparison to Merlo and Stevenson's (2001a) results, we see that the English data extracted from the English HKLaws corpus alone was not very useful (see Chapters 4 and 5). Apart from the inaccuracies of our (relaxed) extraction technique, using a relatively small corpus has the following two problems: first, there might be a data sparseness problem; second, we were only able to find a small test set of 32 verbs. Further, its legal nature may also be problematic. For instance, a corpus of legal documents may contain a higher frequency of passive constructions than, say, a balanced corpus. Roland and Jurafsky (1998) and Roland et al. (2000) also noted that corpus choice can affect the proportion of usage of (polysemous) verbs, and hence the subcategorization frequencies. In light of these observations, we believe that at least for the English feature extraction step, we can benefit from using a larger and more balanced corpus.

7.2 Selection of "L1 Features"

7.2.1 Contributions

In this thesis, we have succeeded in identifying a set of non-English features, for which the statistical distributions over a parallel corpus provide a multi-way distinction of English verbs (see Chapter 3). It is regrettable that we do not have the Chinese equivalent of Levin's (1993) verb classification, which could provide some important syntactic and semantic information to aid our feature selection process. However, the same type of

contrastive linguistic work used in SLA studies is useful for us: given each verb from our target verb classes, our approach is to compare sentence constructions extracted from a bitext. This method works here because:

- Our L1 choice, Chinese, unlike many other non-Indo-European languages, shares many similarities with English. For example, every Chinese sentence must contain a verb and many English verbs have the same part-of-speech as their translated counterpart in Chinese. On the other hand, unlike English and Chinese, some agglutinative languages collapse nouns and verbs into one single part-of-speech. The comparison of verb behaviour between English and, say, Turkish or Eskimo translation equivalents may not be as straightforward.

- Our corpus contains strictly legal documents. The translated documents must be closely aligned with the original version to preserve meaning as much as possible. Sentence by sentence comparison was made easy for us.

It is true that our feature selection step was made easier by the above two conditions. Despite that, we believe our method is extendible to the comparison of any sentence pair using translated verbs. Our feature selection method is novel in that potentially useful features can be identified by observing the similarities in the morpho-syntactic and syntactico-semantic properties across the translations of two languages. Our results show that we have succeeded in selecting useful L1 features.

7.2.2 Future Work: Chinese Verb Classes and Interlingual Representations

In our feature selection step, Chinese features were chosen by comparing English and Chinese linguistic properties such as the different passive sentence constructions. Although this work was inspired by studies in SLA, we have neglected to observe how transfer

effects in humans can help us identify useful features. That is, in many human studies in SLA, there are features responsible for positive transfer, but we have not factored this in as part of our feature selection process. We believe that if a feature is responsible for positive transfer, it is likely to be useful in automatic learning and we should use this as one of our considerations for selecting features.

One kind of feature, Chinese argument structure information, is often discussed in the contrastive linguistic portion of many SLA papers (e.g., Balcom, 1997; Inagaki, 1997; Ju, 2000; Juffs, 1998; Montrul, 2000; Oshita, 2000; Yuan, 1999), but we were hesitant to use it. The main reason is that we did not have access to a Chinese parser while this work was in progress. Another type of information, similar to Chinese argument structure information, is Chinese alternation and verb class information. We are not aware of the existence of such a classification. However, we are unsure of its usefulness, mainly because we do not believe a syntactic and semantic clustering of Chinese verbs would fit into Levin's alternations and verb classes (Levin, 1993). Recall that our goal is to observe possible performance gain using multilingual data. If the clustering of Chinese alternations and verb classes does not "sufficiently overlap" the English clustering, we do not foresee any performance benefits using this type of information. Clearly, we do not know if the same granularity as shown in (Levin, 1993) is appropriate for Chinese verbs until we attempt an equivalent task in Chinese.

Aside from the construction of Chinese alternations and verb classes, we could use an interlingual thematic representation. Olsen et al.(1996; 1998; 2000) use Lexical Conceptual Structure (LCS), a type of interlingual representation, for their machine translation task, in which verb senses are paired with the corresponding argument structures. In any language, an interlingual representation should provide a way of coarsely classifying verbs based on their thematic relations. Rather than relying on a verb classification in a specific language, an interlingual representation can be used as the target classification instead. Assuming that LCS or some other interlingual representation is truly language

neutral and fully developed to support the expressiveness of other languages, clearly we should be able to perform verb classification experiments for other languages.[1] We have already seen one example of automatic verb classification in Japanese using LCS (Oishi and Matsumoto, 1997). Using multilingual features, we can test the generality of interlingual transfer in automatic learning.

7.3 "L1 Transfer" in Automatic Learning

7.3.1 Contributions

Although the nature of the organization of the L1 and L2 lexicon in a second language learner is not well-understood, "L1 transfer" is a well-known phenomenon in the study of SLA. For our work, we were inspired to augment English data with non-English data. Although some researchers have hinted at the possibility of using multilingual data (see section 2.2) in automatic lexical acquisition, it is surprising that there is not more of such work in the area of natural language learning (NLL). This thesis has shown that in NLL, similar to the idea of positive transfer in SLA, if there is a sufficient "overlap"[2] between the two languages, we can benefit from combining the English and non-English data in learning English lexical information.

7.3.2 Future Work – Bi-directional Learning

The motivation of this work is to treat non-English data as an aid to automatic learning of English verb classes, and this notion is manifested as pairing English and non-English data in the actual experiments. However, nothing should keep us from using the same

[1] Although it has been noted by Olsen et al. (2000) and Viegas (2000) that LCS is not language neutral, Olsen et al. (2000) suggested that LCS can be modified to be applicable to the languages of interest.

[2] For a specific class of verbs, there is an overlap between two languages if the translated equivalents are sufficiently "similar" in their usage, e.g., English and Chinese adjectival construction in change-of-state verbs.

pairing for the automatic learning of non-English verb classes. Obviously, in order for this to work for, say, Chinese verbs, the relationship between Chinese alternations and verb classes needs to be determined. One possible area of future work is to explore the type of regularities in Chinese verb syntactic patterns based on semantic class membership along the lines of Levin's (1993) verb classes and alternations.

That said, there is some existing work on classifying Chinese verbs. For example, we have used a type of Chinese verb classification in our work – the CKIP verb classification, which provides a distinction between activity and stative verbs. The state-action dichotomy, though useful for our application, may be too simplistic a classification of Chinese verbs. Other semantic classes, such as finer-grained event class distinctions, are discussed by Dorr and Olsen (1996) and Oishi and Matsumoto (1997). In addition to the CKIP verb classification, others have tried using subcategorization frames as well as thematic role assignments as classification criteria (Her, 1990; Tsao, 1996). There are also discussions on distinguishing Chinese unaccusatives from unergatives (Yuan, 1999). These papers certainly hint at the possibility of constructing a Chinese equivalent to Levin's alternations and verb classes (Levin, 1993).

Chinese verb classification appears to be a long term project and the possibility of bi-directional learning will depend on it. In the short term, as hinted in the previous section, some interlingual representation can be useful in bi-directional learning. To end the thesis on a positive note, we think that Chinese verb classes could be gradually constructed by learning from the intermediate steps of designing a language-neutral representation and performing automatic bi-directional experiments. That is, it is possible that the (unsupervised) learning of clusters of Chinese verbs (as in (Schulte im Walde, 2000) or (Merlo and Stevenson, 2001b), for English verbs) could bootstrap the process of constructing a (linguistically motivated) Chinese verb classification.

Appendix A

Chinese HKLaws Data

This appendix contains the Chinese HKLaws data (the overall frequency and relative feature frequencies of each verb) we used to train our classifier.

A.1 Aligned Method

The following data was collected using the aligned method.

Unaccusative Verbs (alter–decrease), Aligned Method								
Verb	*alter*	*change*	*clear*	*close*	*compress*	*contract*	*cool*	*decrease*
Chi. Freq.	312	96	14	524	110	78	10	14
VA Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VAC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.00	0.00	0.00	0.05	0.00	0.04	0.00	0.00
VC Tag	1.00	0.95	1.00	0.77	1.00	0.63	0.00	0.64
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
VF Tag	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00
VG Tag	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
VH Tag	0.00	0.00	0.00	0.09	0.00	0.05	0.50	0.00
VHC Tag	0.00	0.01	0.00	0.03	0.00	0.00	0.50	0.36
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00

Unaccusative Verbs (alter–decrease), Aligned Method (Cont.)								
Verb	*alter*	*change*	*clear*	*close*	*compress*	*contract*	*cool*	*decrease*
UPenn VA Tag	0.00	0.01	0.00	0.12	0.00	0.20	1.00	0.36
UPenn VV Tag	1.00	0.99	1.00	0.88	1.00	0.80	0.00	0.64
V-V Morph.	0.96	1.00	0.56	0.25	1.00	0.42	0.43	0.50
V-N Morph.	0.04	0.00	0.00	0.01	0.00	0.18	0.00	0.00
V-A Morph.	0.00	0.00	0.00	0.21	0.00	0.00	0.07	0.50
N-V Morph..	0.00	0.00	0.12	0.02	0.00	0.28	0.00	0.00
N-N Morph.	0.00	0.00	0.00	0.01	0.00	0.12	0.00	0.00
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.00	0.00	0.32	0.26	0.00	0.00	0.43	0.00
A-N Morph.	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
A-A Morph.	0.00	0.00	0.00	0.22	0.00	0.00	0.07	0.00
Pass. Part.	0.06	0.00	0.14	0.01	0.00	0.00	0.00	0.00
Peri. Part.	0.02	0.00	0.07	0.03	0.00	0.00	0.00	0.07
Sem. Spec.	0.05	0.01	0.00	0.00	0.00	0.20	0.00	0.00
Avg. Leng.	2.00	1.99	2.00	2.00	2.00	2.00	1.60	2.00

Unaccusative Verbs (diminish–reproduce), Aligned Method								
Verb	*diminish*	*dissolve*	*divide*	*drain*	*flood*	*multiply*	*open*	*reproduce*
Chi. Freq.	18	152	221	20	23	43	247	62
VA Tag	0.00	0.00	0.00	0.57	0.00	0.00	0.00	0.34
VAC Tag	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.05	0.00	0.00	0.00	1.00	0.00	0.00	0.00
VC Tag	0.21	0.23	0.10	0.38	0.00	0.07	0.79	0.50
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VF Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VG Tag	0.00	0.00	0.47	0.00	0.00	0.00	0.00	0.00
VH Tag	0.10	0.00	0.23	0.05	0.00	0.00	0.05	0.00
VHC Tag	0.53	0.77	0.05	0.00	0.00	0.00	0.07	0.16
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.05	0.00	0.00	0.00	0.00	0.93	0.00	0.00
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00

Unaccusative Verbs (diminish–reproduce), Aligned Method (Cont.)								
Verb	*diminish*	*dissolve*	*divide*	*drain*	*flood*	*multiply*	*open*	*reproduce*
UPenn VA Tag	0.68	0.77	0.28	0.05	0.00	0.93	0.13	0.16
UPenn VV Tag	0.32	0.23	0.72	0.95	1.00	0.07	0.87	0.84
V-V Morph.	0.35	0.61	0.47	0.35	0.00	1.00	0.42	1.00
V-N Morph.	0.13	0.00	0.02	0.60	1.00	0.00	0.07	0.00
V-A Morph.	0.52	0.39	0.06	0.05	0.00	0.00	0.04	0.00
N-V Morph..	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.00	0.00	0.40	0.00	0.00	0.00	0.39	0.00
A-N Morph.	0.00	0.00	0.02	0.00	0.00	0.00	0.07	0.00
A-A Morph.	0.00	0.00	0.03	0.00	0.00	0.00	0.01	0.00
Pass. Part.	0.00	0.05	0.01	0.00	0.13	0.00	0.01	0.00
Peri. Part.	0.00	0.00	0.04	0.00	0.00	0.02	0.03	0.00
Sem. Spec.	0.17	0.00	0.00	0.65	0.00	0.00	0.00	0.00
Avg. Leng.	2.00	2.00	2.00	2.00	1.78	1.93	1.77	2.00

Object-Drop Verbs (build–pack), Aligned Method								
Verb	*build*	*clean*	*compose*	*direct*	*hammer*	*knit*	*organise*	*pack*
Chi. Freq.	178	65	37	599	1	5	6	58
VA Tag	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
VAC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VC Tag	0.96	0.09	0.00	0.75	1.00	0.80	1.00	1.00
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00
VF Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VG Tag	0.04	0.00	1.00	0.00	0.00	0.00	0.00	0.00
VH Tag	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00
VHC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00

Object-Drop Verbs (build–pack), Aligned Method (Cont.)								
Verb	*build*	*clean*	*compose*	*direct*	*hammer*	*knit*	*organise*	*pack*
UPenn VA Tag	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00
UPenn VV Tag	1.00	0.09	1.00	1.00	1.00	1.00	1.00	1.00
V-V Morph.	0.51	0.25	1.00	1.00	0.00	1.00	1.00	0.30
V-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20
V-A Morph.	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00
N-V Morph..	0.10	0.00	0.00	0.00	1.00	0.00	0.00	0.30
N-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.20	0.25	0.00	0.00	0.00	0.00	0.00	0.00
A-N Morph.	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-A Morph.	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Pass. Part.	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Peri. Part.	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00
Sem. Spec.	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg. Leng.	1.91	2.00	2.00	2.00	2.00	1.20	2.00	2.00

Object-Drop Verbs (paint–wash), Aligned Method								
Verb	*paint*	*perform*	*play*	*produce*	*recite*	*stitch*	*type*	*wash*
Chi. Freq.	2	369	10	722	38	2	6	43
VA Tag	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
VAC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.02
VC Tag	1.00	1.00	1.00	0.37	0.26	0.50	0.00	0.98
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.00	0.51	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.00	0.74	0.00	0.00	0.00
VF Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VG Tag	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
VH Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VHC Tag	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Object-Drop Verbs (paint–wash), Aligned Method (Cont.)								
Verb	*paint*	*perform*	*play*	*produce*	*recite*	*stitch*	*type*	*wash*
UPenn VA Tag	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00
UPenn VV Tag	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00
V-V Morph.	0.50	1.00	1.00	1.00	0.91	1.00	0.00	0.48
V-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
V-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
N-V Morph..	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.40
A-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pass. Part.	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Peri. Part.	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00
Sem. Spec.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg. Leng.	2.00	2.00	2.00	2.00	1.29	2.00	2.00	1.93

A.2 Unaligned Method

The following data was collected using the unaligned method.

Unaccusative Verbs (alter–decrease), Unaligned Method								
Verb	*alter*	*change*	*clear*	*close*	*compress*	*contract*	*cool*	*decrease*
Chi. Freq.	4158	4181	622	4231	136	9151	126	963
VA Tag	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
VAC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
VC Tag	1.00	0.94	1.00	0.60	1.00	0.63	0.00	0.67
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VF Tag	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00
VG Tag	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
VH Tag	0.00	0.01	0.00	0.04	0.00	0.28	0.52	0.00
VHC Tag	0.00	0.02	0.00	0.11	0.00	0.00	0.48	0.33
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00

Unaccusative Verbs (alter–decrease), Unaligned Method (Cont.)								
Verb	*alter*	*change*	*clear*	*close*	*compress*	*contract*	*cool*	*decrease*
UPenn VA Tag	0.00	0.04	0.00	0.15	0.00	0.29	1.00	0.33
UPenn VV Tag	1.00	0.96	1.00	0.85	1.00	0.71	0.00	0.67
V-V Morph.	0.96	1.00	0.56	0.29	1.00	0.52	0.48	0.60
V-N Morph.	0.04	0.00	0.00	0.03	0.00	0.02	0.00	0.00
V-A Morph.	0.00	0.00	0.00	0.13	0.00	0.00	0.01	0.40
N-V Morph..	0.00	0.00	0.30	0.09	0.00	0.45	0.00	0.00
N-N Morph.	0.00	0.00	0.00	0.03	0.00	0.01	0.00	0.00
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.00	0.00	0.14	0.25	0.00	0.00	0.48	0.00
A-N Morph.	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00
A-A Morph.	0.00	0.00	0.00	0.14	0.00	0.00	0.01	0.00
Pass. Part.	0.01	0.01	0.06	0.01	0.00	0.00	0.00	0.01
Peri. Part.	0.02	0.02	0.03	0.01	0.00	0.00	0.04	0.01
Sem. Spec.	0.05	0.00	0.00	0.00	0.00	0.11	0.00	0.00
Avg. Leng.	2.00	1.96	2.00	2.00	2.00	2.00	1.50	2.00

Unaccusative Verbs (diminish–reproduce), Unaligned Method								
Verb	*diminish*	*dissolve*	*divide*	*drain*	*flood*	*multiply*	*open*	*reproduce*
Chi. Freq.	768	833	1991	1271	197	397	5171	1101
VA Tag	0.00	0.00	0.00	0.42	0.00	0.00	0.00	0.08
VAC Tag	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.02	0.00	0.00	0.00	0.94	0.03	0.00	0.00
VC Tag	0.28	0.70	0.12	0.57	0.06	0.75	0.52	0.84
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06
VF Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VG Tag	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00
VH Tag	0.01	0.00	0.42	0.00	0.00	0.01	0.15	0.00
VHC Tag	0.42	0.30	0.10	0.00	0.00	0.00	0.16	0.01
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.01
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00

Unaccusative Verbs (diminish–reproduce), Unaligned Method (Cont.)								
Verb	*diminish*	*dissolve*	*divide*	*drain*	*flood*	*multiply*	*open*	*reproduce*
UPenn VA Tag	0.44	0.30	0.52	0.00	0.00	0.22	0.31	0.02
UPenn VV Tag	0.56	0.70	0.48	1.00	1.00	0.78	0.69	0.98
V-V Morph.	0.46	0.78	0.43	0.57	0.00	1.00	0.29	1.00
V-N Morph.	0.02	0.00	0.01	0.43	0.80	0.00	0.15	0.00
V-A Morph.	0.52	0.22	0.12	0.00	0.10	0.00	0.06	0.00
N-V Morph..	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.00	0.00	0.36	0.00	0.00	0.00	0.28	0.00
A-N Morph.	0.00	0.00	0.01	0.00	0.00	0.00	0.15	0.00
A-A Morph.	0.00	0.00	0.07	0.00	0.10	0.00	0.06	0.00
Pass. Part.	0.01	0.02	0.01	0.00	0.10	0.00	0.00	0.00
Peri. Part.	0.02	0.02	0.03	0.02	0.01	0.01	0.01	0.02
Sem. Spec.	0.03	0.00	0.00	0.43	0.00	0.00	0.00	0.00
Avg. Leng.	2.00	2.00	2.00	2.00	1.51	1.21	1.58	2.00

Object-Drop Verbs (build–pack), Unaligned Method								
Verb	*build*	*clean*	*compose*	*direct*	*hammer*	*knit*	*organise*	*pack*
Chi. Freq.	4682	627	1539	20444	4	66	1537	467
VA Tag	0.00	0.00	0.00	0.00	0.00	0.39	0.00	0.00
VAC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VC Tag	0.96	0.08	0.00	0.90	1.00	0.61	1.00	1.00
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00
VF Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VG Tag	0.04	0.00	1.00	0.00	0.00	0.00	0.00	0.00
VH Tag	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00
VHC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00

Object-Drop Verbs (build–pack), Unaligned Method (Cont.)								
Verb	*build*	*clean*	*compose*	*direct*	*hammer*	*knit*	*organise*	*pack*
UPenn VA Tag	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.00
UPenn VV Tag	1.00	0.08	1.00	1.00	1.00	1.00	1.00	1.00
V-V Morph.	0.44	0.23	1.00	1.00	0.00	1.00	1.00	0.38
V-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
V-A Morph.	0.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00
N-V Morph..	0.15	0.00	0.00	0.00	1.00	0.00	0.00	0.38
N-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.20	0.23	0.00	0.00	0.00	0.00	0.00	0.00
A-N Morph.	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-A Morph.	0.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00
Pass. Part.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Peri. Part.	0.00	0.04	0.00	0.00	0.00	0.00	0.01	0.03
Sem. Spec.	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg. Leng.	1.96	2.00	2.00	2.00	2.00	1.39	2.00	2.00

Object-Drop Verbs (paint–wash), Unaligned Method								
Verb	*paint*	*perform*	*play*	*produce*	*recite*	*stitch*	*type*	*wash*
Chi. Freq.	4	16315	1114	15374	11205	3	17	328
VA Tag	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.08
VAC Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VB Tag	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00
VC Tag	1.00	1.00	1.00	0.16	0.07	0.67	0.00	0.91
VCL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VD Tag	0.00	0.00	0.00	0.74	0.00	0.00	0.00	0.00
VE Tag	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00
VF Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VG Tag	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
VH Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
VHC Tag	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00
VI Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VJ Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VK Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VL Tag	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Object-Drop Verbs (paint–wash), Unaligned Method								
Verb	*paint*	*perform*	*play*	*produce*	*recite*	*stitch*	*type*	*wash*
UPenn VA Tag	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.01
UPenn VV Tag	1.00	1.00	1.00	0.92	1.00	1.00	1.00	0.99
V-V Morph.	0.50	1.00	1.00	1.00	0.80	1.00	0.00	0.58
V-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
V-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
N-V Morph.	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-V Morph.	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.40
A-N Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-A Morph.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pass. Part.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Peri. Part.	0.00	0.00	0.01	0.02	0.01	0.00	0.00	0.01
Sem. Spec.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Appendix B

English HKLaws Data

This appendix contains the English HKLaws data (the overall frequency and relative feature frequencies of each verb) we used to train our classifier.

Unaccusative Verbs (alter–decrease)								
Verb	*alter*	*change*	*clear*	*close*	*compress*	*contract*	*cool*	*decrease*
Eng. Freq.	145	59	7	311	10	80	10	12
Transitivity	0.88	0.68	1.00	0.81	0.30	0.42	0.70	0.92
Passive Voice	0.94	0.65	1.00	0.90	1.00	0.49	1.00	1.00
VBN Tag	0.98	0.94	0.86	0.95	1.00	0.77	1.00	1.00
Causativity	0.01	0.07	0.00	0.08	0.00	0.16	0.20	0.00
Animacy	0.12	0.00	0.00	0.03	0.00	0.28	0.00	0.00

Unaccusative Verbs (diminish–reproduce)								
Verb	*diminish*	*dissolve*	*divide*	*drain*	*flood*	*multiply*	*open*	*reproduce*
Eng. Freq.	13	89	190	8	12	40	153	18
Transitivity	0.77	0.91	0.86	1.00	0.92	0.53	0.74	0.78
Passive Voice	0.92	0.98	1.00	1.00	1.00	1.00	0.91	0.89
VBN Tag	1.00	0.99	1.00	0.93	1.00	1.00	0.93	0.95
Causativity	0.15	0.17	0.03	0.00	0.25	0.07	0.12	0.06
Animacy	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00

Object-Drop Verbs (build–pack)								
Verb	*build*	*clean*	*compose*	*direct*	*hammer*	*knit*	*organise*	*pack*
Eng. Freq.	288	21	63	421	1	3	12	52
Transitivity	0.54	0.95	0.68	0.68	1.00	0.00	0.50	0.73
Passive Voice	1.00	0.95	1.00	0.79	1.00	1.00	1.00	1.00
VBN Tag	0.99	0.89	1.00	0.92	1.00	1.00	1.00	0.99
Causativity	0.43	0.00	0.05	0.15	0.00	0.00	0.50	0.25
Animacy	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00

Object-Drop Verbs (paint–wash)								
Verb	*paint*	*perform*	*play*	*produce*	*recite*	*stitch*	*type*	*wash*
Eng. Freq.	29	337	18	644	20	1	3	17
Transitivity	0.90	0.65	0.89	0.69	0.65	1.00	0.33	0.82
Passive Voice	1.00	0.90	0.86	0.94	1.00	1.00	1.00	0.82
VBN Tag	1.00	0.96	1.00	0.97	1.00	1.00	1.00	0.80
Causativity	0.03	0.09	0.00	0.13	0.10	0.00	0.00	0.00
Animacy	0.00	0.03	0.00	0.01	0.00	0.00	0.00	0.00

Appendix C

WSJ Data

This appendix contains the WSJ data (the overall frequency and relative feature frequencies of each verb) we used to train our classifier.

Unaccusative Verbs (alter–decrease)								
Verb	*alter*	*change*	*clear*	*close*	*compress*	*contract*	*cool*	*decrease*
Eng. Freq.	521	5871	1123	23786	61	467	215	564
Transitivity	0.52	0.45	0.58	0.20	0.28	0.32	0.33	0.43
Passive Voice	0.14	0.12	0.17	0.05	0.18	0.06	0.10	0.02
VBN Tag	0.82	0.75	0.59	0.11	0.93	0.42	0.86	0.30
Causativity	0.06	0.43	0.22	0.36	0.00	0.19	0.14	0.14
Animacy	0.15	0.07	0.04	0.01	0.00	0.10	0.04	0.01

Unaccusative Verbs (diminish–reproduce)								
Verb	*diminish*	*dissolve*	*divide*	*drain*	*flood*	*multiply*	*open*	*reproduce*
Eng. Freq.	445	211	1489	145	226	100	3702	47
Transitivity	0.26	0.54	0.52	0.57	0.60	0.00	0.44	0.00
Passive Voice	0.11	0.28	0.45	0.22	0.28	0.00	0.04	0.00
VBN Tag	0.92	0.72	0.94	0.76	0.76	0.85	0.18	0.79
Causativity	0.08	0.12	0.09	0.09	0.05	0.00	0.40	0.00
Animacy	0.02	0.07	0.12	0.02	0.05	0.00	0.06	0.00

Object-Drop Verbs (build–pack)								
Verb	*build*	*clean*	*compose*	*direct*	*hammer*	*knit*	*organise*	*pack*
Eng. Freq.	4137	149	370	1129	247	31	1417	360
Transitivity	0.40	0.33	0.33	0.48	0.40	0.03	0.00	0.34
Passive Voice	0.20	0.19	0.29	0.15	0.31	0.00	0.00	0.19
VBN Tag	0.79	0.79	0.86	0.67	0.89	0.97	0.85	0.81
Causativity	0.20	0.08	0.03	0.17	0.06	0.00	0.00	0.02
Animacy	0.19	0.35	0.23	0.16	0.11	0.00	0.00	0.30

Object-Drop Verbs (paint–wash)								
Verb	*paint*	*perform*	*play*	*produce*	*recite*	*stitch*	*type*	*wash*
Eng. Freq.	467	1042	2593	4076	32	31	55	103
Transitivity	0.38	0.26	0.54	0.42	0.72	0.26	0.00	0.35
Passive Voice	0.14	0.14	0.06	0.12	0.00	0.26	0.00	0.21
VBN Tag	0.72	0.80	0.38	0.72	0.00	1.00	0.80	0.82
Causativity	0.07	0.08	0.40	0.31	0.00	0.00	0.00	0.00
Animacy	0.28	0.18	0.21	0.10	0.39	0.00	0.00	0.17

Bibliography

Aone, C. and McKee, D. (1996). Acquiring predicate-argument mapping information in multilingual texts. In Boguraev, B. and Pustejovsky, J., editors, *Corpus Processing for Lexical Acquisition*, pages 191–202. MIT Press.

Balcom, P. (1997). Why is this happened? Passive morphology and unaccusativity. *Second Language Research*, 13(1):1–9.

Brill, E. (1993). Transformation-based error-driven parsing. In *Proceedings of the Third International Workshop on Parsing Technologies*, Tilburg, The Netherlands.

Chang, C. H.-H. (1990). On serial verbs in Mandarin Chinese: VV compounds and co-verbial phrases. In Joseph, B. D. and Zwicky, A. M., editors, *When Verbs Collide: Papers from the 1990 Ohio State Mini-Conference on Serial Verbs*, pages 288–315. The Ohio State University, Columbus, Ohio.

Chang, C. H.-H. (1998). V-V compounds in Mandarin Chinese: Argument structure and semantics. In Packard, J. L., editor, *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*, pages 77–101. Mouton de Gruyter, New York.

Chang, L.-L., Chen, K.-J., and Huang, C.-R. (1999). Alternation across semantic fields: A study of Mandarin verbs of emotion. In *Proceedings of The 13th Pacific Asia Conference on Language, Information and Computation*, pages 39–50, Taipei, Taiwan.

Chen, K.-J. and Hong, W.-M. (1996). 中文裡『動–名』述賓結構與『動–名』偏正結構的分析 (Argument structures of Chinese verb-noun compounds). *Communications of Chinese and Oriental Languages Information Processing Society (COLIPS)*, 6(2):73–79.

Dorr, B. and Olsen, M. (1996). Multilingual generation: The role of telicity in lexical choice and syntactic realization. *Machine Translation*, 11:37–74.

Fung, P. (1998). A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In Farwell, D., Gerber, L., and Hovy, E., editors, *Third Conference of the Association for Machine Translation in the Americas*, pages 1–16. Springer.

Fung, P. and Lo, Y. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 414–420, Montreal, Canada.

Fung, P. and McKeown, K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.

Gale, W. and Church, K. (1991). A program for aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California.

Helms-Park, R. (1997). *Building an L2 Lexicon: The Acquisition of Verb Classes Relevant to Causativization in English by Speakers of Hindi-Urdu and Vietnamese.* PhD thesis, University of Toronto, Toronto, Canada.

Her, O. S. (1990). *Grammatical Functions and Verb Subcategorization in Mandarin Chinese.* Crane Publishing Co., Taipei, Taiwan.

Her, O. S. (1996). Variation of the VO construction in Chinese: a synchronic account. *Linguistics*, 34:733–751.

Her, O. S. (1997). Interaction and explanation: The case of variation in Chinese VO construction. *Journal of Chinese Linguistics*, 25(1):146–165.

Hsu, L. and Wu, Z. (1994). Translating English change-of-state verbs into Chinese serial verb compounds. *Computer Processing of Chinese and Oriental Languages*, 8. Vol. 8 Supplement.

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 Chinese Language Processing Workshop*, Hong Kong.

Huang, S. (1998). Chinese as a headless language in compounding morphology. In Packard, J. L., editor, *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*, pages 261–283. Mouton de Gruyter, New York.

Ide, N. (1999). Parallel translations as sense discriminators. In *Proceedings of SIGLEX 1999: Standardizing Lexical Resources*, pages 52–61, College Park, MD.

Ide, N. (2000). Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34:223–234.

Inagaki, S. (1997). Japanese and Chinese learners' acquisition of the narrow-range rules for the dative alternation in English. *Language Learning*, 47(4):637–669.

Ju, M. K. (2000). Overpassivization errors by second language learners. *Studies in Second Language Acquisition*, 22:85–111.

Juffs, A. (1998). Some effects of first language argument structure and morphosyntax on second language sentence processing. *Second Language Research*, 14(4):406–424.

Juffs, A. (2000). An overview of the second language acquisition of links between verb semantics and morpho-syntax. In Archibald, J., editor, *Second Language Acquisition and Linguistic Theory*, pages 170–179. Blackwell Publishers.

Korhonen, A.-L. (1997). Acquiring subcategorisation from textual corpora. Master's thesis, University of Cambridge, Cambridge, UK.

Lapata, M. and Brew, C. (1999). Using subcategorization to resolve verb class ambiguity. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP) and Very Large Corpora*, pages 266–274, College Park, MD.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Levin, B. and Rappaport Hovav, M. (1995). *Unaccusativity at the Syntax-Lexical Semantics Interface*. The MIT Press.

Lin, C., Wei, W., and Huang, J. (1997). 漢語的類雙賓動詞現象 (Chinese pseudo-ditransitive verbs). In *Proceedings of The 9th North American Conference on Chinese Linguistics (NACCL-9)*, Victoria, B.C., Canada.

Liu, S., Chen, K., Chang, L., and Chin, Y. (1995). Automatic part-of-speech tagging for Chinese corpora. *Computer Processing of Chinese and Oriental Languages*, pages 31–48.

Lua, K. (1997). An efficient inductive unsupervised semantic tagger. *Computer Processing of Chinese and Oriental Languages*, 11(1).

McCarthy, D. (2000). Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of Applied Natural Language Processing and North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, pages 256–263, Seattle, WA.

McCarthy, D. and Korhonen, A.-L. (1998). Detecting verbal participation in diathesis alternations. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 1493–1495, Montreal, Canada.

Melamed, I. D. and Marcus, M. P. (1998). Automatic construction of Chinese-English translation lexicons. Technical Report 98-28, University of Pennsylvania, Philadelphia, PA.

Merlo, P. and Stevenson, S. (2001a). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27(3):393–408.

Merlo, P. and Stevenson, S. (2001b). Unsupervised learning of verb classes from lexical statistics. In *Proceedings of the 7th Annual Conference on Architectures and Mechanics for Language Processing (AMLaP-2001)*, Saarbrücken, Germany. To appear.

Montrul, S. (2000). Transitivity alternations in L2 acquisition – towards a modular view of transfer. *Studies in Second Language Acquisition*, 22(2):229–273.

Oishi, A. and Matsumoto, Y. (1997). Detecting the organization of semantic subclasses of Japanese verbs. *International Journal of Corpus Linguistics*, 2(1):65–89.

Olsen, M. (1998). Translating English and Mandarin verbs with argument structure (mis)matches using LCS representation. In *Proceedings of the Second SIG-IL Workshop*, Philadelphia, PA.

Olsen, M., Traum, D., Ess-Dykema, C. V., Weinberg, A., and Dolan, R. (2000). Telicity as a cue to discourse structure in Chinese-English machine translation. In *Proceedings of the Third SIG-IL Workshop*, Seattle, Washington.

Oshita, H. (2000). *What is happened* may not be what appears to be happening: a corpus study 'passive' unaccusatives in L2 English. *Second Language Research*, 16(4):293–324.

Palmer, M. and Wu, Z. (1995). Verb semantics for English-Chinese translation. *Machine Translation*, pages 1–32.

Pao, S. S. (2000). *Sentence and Word Alignment Between Chinese and English*. PhD thesis, University of Sheffield, Sheffield, UK.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of The Empirical Methods in Natural Language Processing Conference*, Philadelphia, PA.

Resnik, P. and Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of SIGLEX 1997: Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC.

Resnik, P. and Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Roland, D. and Jurafsky, D. (1998). How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Cnference on Computational Linguistics (COLING-ACL 1998)*.

Roland, D., Jurafsky, D., Menn, L., Gahl, S., Elder, E., and Riddoch, C. (2000). Verb subcategorization frequency differences between business-news and balanced corpora: The role of verb sense. In *Proceedings of the Association for Computational Linguistics (ACL 2000) Workshop on Comparing Corpora*, Hong Kong.

Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation

behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 747–753, Saarbrücken, Germany.

Siegel, E. V. and McKeown, K. R. (2000). Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628.

Starosta, S., Kuiper, K., Ng, S.-A., and Wu, Z.-Q. (1998). On defining the Chinese compound word: Headedness in Chinese compounding and Chinese VR compounds. In Packard, J. L., editor, *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*, pages 347–370. Mouton de Gruyter, New York.

Stevenson, S. and Merlo, P. (1997). Lexical structure and parsing complexity. *Language and Cognitive Processes*, 12(2/3):349–399.

Thompson, S. A. (1973). Transitivity and some problems with the BA-Construction in Mandarin Chinese. *Journal of Chinese Linguistics*, 1(2).

Trask, R. L. (1993). *A Dictionary of Grammatical Terms in Linguistics*. Routledge Inc.

Tsao, F. (1996). On verb classification in Chinese. *Journal of Chinese Linguistics*, 24(1):138–191.

Verspoor, C. M. (1997). *Contextually-Dependent Lexical Semantics*. PhD thesis, University of Edinburgh, Edinburgh, UK.

Viegas, E. (2000). Critique – telicity as a cue to discourse structure in Chinese-English machine translation by Olsen et al., 2000. *The Third SIG-IL Workshop*. `http://crl.nmsu.edu/Events/FWOI/ThirdWorkshop/FinalPapers/olsen.viegas.html`.

Wang, C. and Lee, T. H.-T. (1999). L2 acquistion of conflation classes of prenominal adjectival participles. *Language Learning*, 49(1):1–36.

Wu, X. (1996). Lexical mapping in the Ba-Construction. Master's thesis, Stanford University, Stanford, CA.

Xia, F. (1999). *Part-of-Speech Tagset Guidelines for Chinese Treebank Project*. Philadelphia, PA, draft edition.

Yuan, B. (1999). Acquiring the unaccusative/unergative distinction in a second language: evidence from English-speaking learners of L2 Chinese. *Linguistics*, 37(2):275–296.

Zhang, M. and Sheng, L. (1997). Tagging Chinese corpus using statistics techniques and rule techniques. In *Proceedings of 1997 International Conference on Computer Processing of Oriental Languages (ICCPOL97)*, pages 503–506, Hong Kong.