International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI Core, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而
治後世聖人易之以書
契百官以治萬民以察
說文敘曰蓋文字者經
藝之本宣教明化之始
前人所以垂後後人所
以識古故曰本立而道
生知天下之至賾而不
可亂也教化既萌文心
雕龍則謂人之立言因
字而生句積句而成章
積章而成篇篇之彪炳

Vol.14 No.3 September 2009 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

The editing of this journal is subsidized by Center for Humanities Research National Science Council in 2009.

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 14, no. 3, September 2009

Contents

Papers

The Association for Computational Linguistics and Chinese Language Processing

Computational Linguistics and Chinese Language Processing

Vol. 14, No. 3, September 2009, pp. 237-256

Modeling Taiwanese POS Tagging Using Statistical Methods and Mandarin Training Data

Un-Gian Iunn[*], Jia-hung Tai[+], Kiat-Gak Lau[#], Cheng-yan Kao[*], and Keh-jiann Chen[+]

Abstract

In this paper, we introduce a POS tagging method for Taiwan Southern Min. We use the more than 62,000 entries of the Taiwanese-Mandarin dictionary and 10 million words of Mandarin training data to tag Taiwanese. The literary written Taiwanese corpora have both Romanized script and Han-Romanization mixed script, and include prose, novels, and dramas. We follow the tagset drawn up by CKIP.

We developed a word alignment checker to assist with the word alignment for the two scripts. It searches the Taiwanese-Mandarin dictionary to find corresponding Mandarin candidate words, selects the most suitable Mandarin word using an HMM probabilistic model from the Mandarin training data, and tags the word using an MEMM classifier.

We achieve an accuracy rate of 91.6% on Taiwanese POS tagging work, and we analyze the errors. We also discover some preliminary Taiwanese training data.

Keywords: Taiwan Southern Min, POS tagging, written Taiwanese, Hidden Markov Model, Maximal Entropy Markov Model.

[*] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
E-mail: {d93001, cykao}@csie.ntu.edu.tw

[+] Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: {glaxy, kchen}@iis.sinica.edu.tw

[#] Independent scholar
E-mail: kiatgak01@gmail.com

1. Introduction

1.1 Background

There are about 46 million Southern Min speakers in the world. If we list languages by the size of their speaking population, Southern Min is ranked 21. The Southern Min speakers are mainly distributed in eight countries(Gordon, 2005). It is an important language that has received very little attention.

The percentage of Southern Min speakers in Taiwan was over 70% (Huang, 1995). Taiwan has the highest percentage of Southern Min speakers in the world. We will call this language as “Taiwanese” for simplification in this paper.

Many different types of written Taiwanese systems exist. Among these systems, the Han character script and one of the Romanized scripts (Pe̍h-ōe-jī, 白話字, *abbrev*. POJ, vernacular writing) are the most popular. Also, the mixture of the above two scripts, called the Han-Romanization mixed script (*abbrev*. as HR mixed script), has been adopted by many people (Iunn, 2009).

1.2 Motivation

In order to establish the bases of written Taiwanese processing, we have constructed some tools over the past few years, including an online Taiwanese syllable dictionary (Iunn, 2003a); an online Taiwanese-Mandarin dictionary (abbrev. OTMD) (Iunn, 2000, 2003b); a 5,800,000 syllable HR mixed script and 3,400,000 syllable POJ script Taiwanese corpus; the online Taiwanese concordancer system based on this corpus (Iunn, 2003c; Iunn & Lau, 2007); preliminary Taiwanese word frequency reports for the Taiwanese POJ and HR mixed scripts, based on the above Taiwanese corpus (Iunn, 2005); the digital archive database for written Taiwanese (*abbrev*. DADWT) literature data with POJ and HR mixed script paragraph alignment (Iunn, 2007); *etc*.

We intend to annotate the Taiwanese corpus with POS markers for more advanced applications, including Taiwanese tone sandhi TTS system improvement (Iunn *et al*. 2007), Taiwanese Treebank construction, *etc*.

1.3 Problem

The primary difficulty encountered in the POS tagging of Taiwanese corpora is the question, “What is the Taiwanese POS tagset?” To date, no standard tagset for Taiwanese has been proposed. Under the circumstances, we have temporarily employed the Chinese POS tagset established by the CKIP Group of Academia Sinica (CKIP, 1993). Unfortunately, we still encountered some problems because we did not have a Taiwanese dictionary that contained the Mandarin POS tagset. The existing Taiwanese dictionaries merely contain basic vocabulary words, that is, nouns, verbs, adjectives, *etc*.

Moreover, there was another problem to surmount – manpower shortage. We did not have enough manpower to fully execute the POS tagging of the Taiwanese corpora.

Therefore, we proposed employing statistical procedures with the existing Mandarin resources and the OTMD to automatically complete the Taiwanese POS tagging. We used the Mandarin language model under the assumption that the word sequence in Taiwanese is similar to Mandarin.

1.4 Review

Shi (2006) translated the Mandarin sentences in the book, "Modern Chinese 800 words '現代漢語八百詞' " (by Shu-xiang Lü) into Taiwanese and Hakka to establish the T3 corpus and developed some editing tools to help in the construction of the T3 Treebank. Chou (2006) used the Brill tagger based on the HMM model to tag words in the T3 Treebank. They used a tagset size of 26, and attained tagging accuracy rates of 92.80% and 85.59% for the training and test data, respectively.

T3 Treebank has not been released publicly. Thus, we decided to use different tagsets and different tagged corpora in our experiments.

2. POS Tagging Method

Figure 1 shows our system architecture diagram.

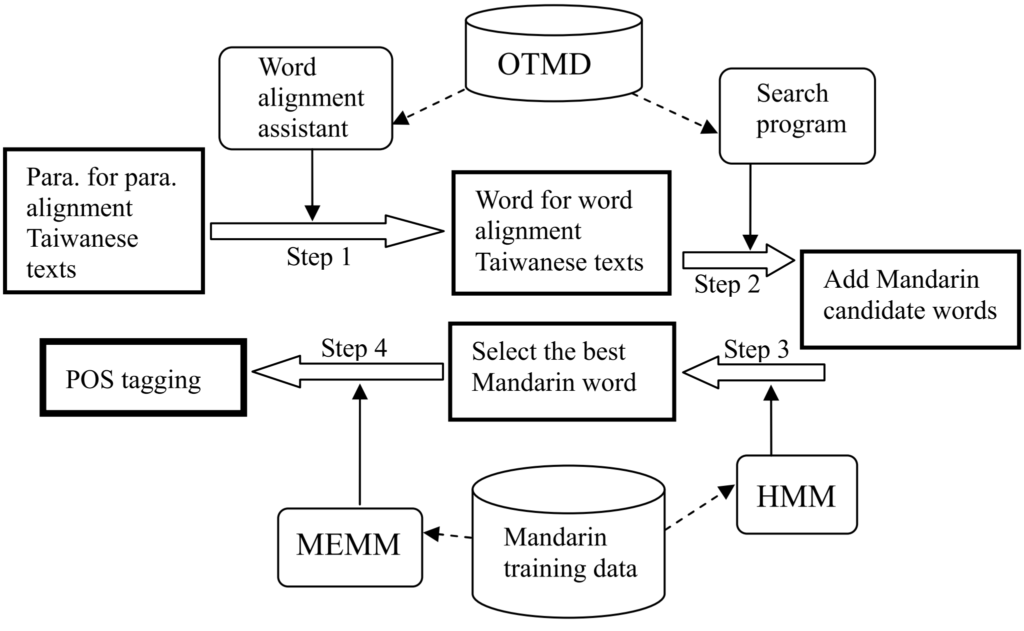

OTMD
Word alignment assistant
Search program
Para. for para. alignment Taiwanese texts
Step 1
Word for word alignment Taiwanese texts
Step 2
Add Mandarin candidate words
Step 3
Select the best Mandarin word
Step 4
POS tagging
HMM
MEMM
Mandarin training data

Figure 1. Taiwanese POS Tagging System Architecture Diagram

At first, the text contains both POJ and HR mixed scripts with paragraph by paragraph alignment. Step 1 converts the texts to word alignment form. Step 2 adds the Mandarin candidate words (translations). Step 3 selects the best Mandarin translation using the HMM model. Finally, we decide the POS tagging of each word using the MEMM model. The following subsection will describe this process in detail.

For example, the original texts are

“Tâi-ôan tē-it kôan ê Gio̍k-san ê hū-kūn khah kē ê só͘-chāi ... ” and
“台灣 第一 懸 ê 玉山 ê 附近 較 低 ê 所在 ... ”
Taiwan first high of Mt.Jade of nearby more low of place

Step 1 converts the texts to word alignment form:

“台灣/Tâi-ôan 第一/tē-it 懸/kôan ê/ê 玉山/Gio̍k-san ê/ê 附近/hū-kūn
較/khah 低/kē ê/ê 所在/só͘-chāi … ”

Then, Step 2 adds the Mandarin translations:

“台灣/Tâi-ôan{台灣} 第一/tē-it{第一;絕頂} 懸/kôan{高} ê/ê{的} 玉山
/Gio̍k-san{玉山}
ê/ê{的} 附近/hū-kūn{附近} 較/khah{較} 低/kē{低} ê/ê{的} 所在/só͘-chāi(去處;
地方;角頭;所在;處所;場所;間量} …”

Step 3 selects the best Mandarin translation using the HMM model (we omit the original Taiwanese texts):

“台灣 第一 高 的 玉山 的 附近 較 低 的 地方 …”

Finally, Step 4 decides the POS tagging of each word using the MEMM model:

“台灣/Tâi-ôan(Nc) 第一/tē-it(Neu) 懸/kôan(VH) ê/ê(DE) 玉山/ Gio̍k -san(Nc)
ê/ê(DE)
附近/hū-kūn(Nc) 較/khah(Dfa) 低/kē(VH) ê/ê(DE) 所在/só͘-chāi(Na)… ”

We will illustrate our work with Figure 1 in the following subsections.

2.1 Origin of the Corpus

The corpus we chose is part of the DADWT project achievements of the National Museum of Taiwan Literature. It contains both POJ and HR mixed scripts with paragraph by paragraph alignment, including novels, prose, dramas, and poems (Iunn, 2007).

2.2 Word by Word Alignment

First, we developed a word alignment program to aid manual processing. We arranged the word alignment of the two scripts, where the paragraphs were already aligned. This program not only collates the number of syllables in the two scripts, but it also compares and contrasts the two scripts with the entries of the OTMD. If the program does not find the two scripts within the same entry, it highlights the corresponding words to remind the user that the word may be an unknown word, an inconsistent usage of the Han character, or a typographical error.

The OTMD was announced and has been online since 2000. The main data provider is Robert L. Cheng, but many anonymous contributors also offer entries and correct the typographical errors. There are a total of more than 62,000 entries. The URL is http://iug.csie.dahan.edu. tw/q. This dictionary offers POJ, HR mixed script, and Mandarin fields, with the POJ field also offering the different accents. The pronunciation function was added in 2006, and English translation was added to more than 10,000 entries in 2007 based on Embree (1984), which contains English, Mandarin, and POJ fields.

2.3 Finding the Corresponding Mandarin Candidate Words

Next, we continued to search for the corresponding Mandarin candidate words from the POJ and HR mixed script word pairs via the OTMD. The mapping was one-to-many. In short, a Taiwanese word pair would have more than one Mandarin word counterpart. For example, "愛/ài" in Taiwanese has the meanings of "愛"'love (person),' "喜歡"'like (thing),' "要" 'want to,' "需要"'need to,' *etc.* in Mandarin. Nevertheless, we were not able to find counterparts for certain words, since they were not contained in the OTMD. We also found some that had different HR mixed script usage.

For instance, the Taiwanese word that appears as "較贏/khah-iâⁿ" 'more than' in the corpus appears as "khah 贏/khah-iâⁿ" in the dictionary. With regard to problems of this nature, we applied the following solution. If the POJ and HR mixed script word pair could not be found, we temporarily removed the HR mixed script and searched for the Mandarin word counterpart again using the POJ script. If the characters of HR mixed script were all Han characters, we regarded the Han characters as one of a Mandarin candidate word (assuming that the word is common to both Taiwanese and Mandarin).

This method might increase the number of the Mandarin candidate words, especially for single syllable words. For instance, the word pair “轉/chōan”‘turn’ appears in the text. We could not find an entry that contains both “轉” and “chōan” in the OTMD. The corresponding Mandarin translations of “chōan” in the dictionary are “扭”’twist’ and “上” ‘up’. We added “轉”‘turn’ as the supplementary Mandarin translation, but the meanings of these three words differ.

Table 1. Partial Entries of the OTMD

HR Mixed Script	POJ Script	Mandarin Translation
chōan	chōan	扭
撰	chōan	上

Note: There exists not “轉/chōan” entry in the OTMD. The Mandarin translation of “轉/chōan” will be “扭,” “上” and “轉”

If the strategy was still unable to find any results, the HR mixed script was directly recognized as the Mandarin candidate word. For instance, no dictionary entry was found for the word pair appearing as “有形/iú-hêng”‘tangible’ in the text, neither could one be found in the search using the POJ script “iú-hêng.” So, the HR mixed script “有形” was directly recognized as the Mandarin candidate word (Lau, 2007).

2.4 Selecting the Best Mandarin Translation

We employed the Hidden Markov Model and Viterbi algorithm, and we made use of the bigram word training data of the ten-million word balanced Sinica corpus of the CKIP Group of Academia Sinica to select the most appropriate corresponding Mandarin word from the Mandarin candidate words. Figure 2 is an example. The selected words are boxed and bold.

Taiwanese Word	對/ Tùi ‘from′	古早/ kó͘ –chá ‘ago’	以來/ í-lâi ‘since’	琴/ khîm ‘instrument’	有/ ū ‘has’	濟濟/ chē-chē ‘many’	款/ khóan ‘appearance’
Corresponding Mandarin Word(s)	**從** w_{11} 對 w_{12} 對子 w_{13} 對於 w_{14}	**以前** w_{21} 古代 w_{22} 古時候 w_{23} 從前 w_{24}	**以來** w_{31}	**琴** w_{41}	**有** w_{51}	濟濟 w_{61} **很多** w_{62}	**樣子** w_{71} 樣式 w_{72} 整理 w_{73}
	$w_1 = w_{11}$	$w_2 = w_{21}$	$w_3 = w_{31}$	$w_4 = w_{41}$	$w_5 = w_{51}$	$w_6 = w_{62}$	$w_7 = w_{71}$

Figure 2. An Example of Selecting the Best Mandarin Translation

Assume that a particular sentence contains *m* words. The first word, w_1, is selected from the candidate words of $w_{11}, w_{12}, ..., w_{1n_1}$; the second word, w_2, is selected from the candidate words of $w_{21}, w_{22}, ..., w_{2n_2}$; and the m^{th} word, w_m, is selected from the candidate words of $w_{m1}, w_{m2}, ..., w_{mn_m}$. $\hat{S} = w_1 w_2 \cdots w_m$, which is the most probable word sequence, is selected from the candidate words, such that P($\hat{S} = w_1 w_2 \cdots w_m$) is maximized.

The HMM assumes that the word w_i is only influenced by the previous word w_{i-1}, thus:

$$P(\hat{S} = w_1 w_2 \cdots w_m) \cong P(w_1) \times \prod_{i=2}^{m} P(w_i \mid w_{i-1}) \tag{1}$$

Therefore, it searches for the word sequence $\hat{S} = w_1 w_2 \cdots w_m$, which maximizes

$$\log P(w_1) + \sum_{i=2}^{m} \log P(w_i \mid w_{i-1}) \tag{2}$$

We use the Laplace smoothing method to solve the problem of $P(w_i \mid w_{i-1}) = 0$, where no bigram of $w_{i-1} w_i$ could be found in the training data in other words. It should be noted that the word string $\hat{S}$ may not be a legal Mandarin sentence.

In practice, we use the Viterbi algorithm to eliminate repeated computation and reduce the time complexity from exponential time to polynomial time. If a sentence S has *m* words, and every word has *n* candidate words, the time complexity will be $O(n^m)$. The Viterbi algorithm reduces the time complexity to $O(n^2 \times m)$ (Manning & Schütze, 1999).

2.5 Selecting the Most Appropriate POS According to the Corresponding Mandarin Word

We applied the Maximal Entropy Markov Model (MEMM) to the POS tag selection.

Manning and Schütze (1999) stated that “Maximum entropy modeling is a framework for integrating information from many heterogeneous information sources for classification. The data for a classification problem is described as a number of features. Each feature corresponds to a constraint on the model. …Choosing the maximum entropy model is motivated by the desire to preserve as much uncertainty as possible.”

MEMM includes a set of possible word and tag contexts, or “histories” (*H*), and the POS tagging set (*T*):

$$p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)} \tag{3}$$

where $h \in H, t \in T$, π is a normalization constant, $\{\mu, \alpha_1, ..., \alpha_k\}$ are the positive model parameters, and $\{f_1, ..., f_k\}$ stands for the features $f_j(h,t) \in \{0,1\}$. Parameter α_j corresponds to the feature f_j. The parameters $\{\mu, \alpha_1, ..., \alpha_k\}$ are then chosen to maximize the likelihood of the training data using p:

$$L(p) = \prod_{i=1}^{n} p(h_i, t_i) = \prod_{i=1}^{n} \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h_i,t_i)} \tag{4}$$

As for the POS tag t_i of the target word w_i, we selected ten features including:

(a) Words – five types of feature patterns: $w_i, w_{i-1}, w_{i-2}w_{i-1}, w_{i+1}, w_{i+1}w_{i+2}$.

(b) POS – two types of feature patterns: $t_{i-1}, t_{i-2}t_{i-1}$.

(c) Morpheme – three types of feature patterns: m_1, m_2, m_n.

The feature patterns m_1, m_2, m_n are designated to manipulate the unknown words. If w_i is an unknown word, we segment w_i with a maximal matching strategy; thus, $w_i = m_1 m_2 \cdots m_n$ and, under certain circumstances, $m_2 = m_3 = \cdots = m_n$. If w_i is a known word, the three morpheme features are set to null. Moreover, if w_i is at the beginning or end of a sentence, certain features are likewise given a null value. For instance, when *i=1*, the feature values of $w_{i-1}, w_{i-2}w_{i-1}, t_{i-1}, t_{i-2}t_{i-1}$, *etc*. are also null (Berger *et al*., 1996; McCallum *et al*., 2000; Rabiner, 1989; Ratnaparkhi, 1996; Samuelsson, 2003; Tai, 2007; Tsai & Chen, 2004).

In MEMM, the dependencies of observations are flexibly modeled whereas HMM assumes that observations are independent. We think MEMM is more suitable for the POS tagging task.

We used the “Maximum Entropy Modeling Toolkit for Python and C++” package provided by Zhang Le to implement our system (Le, 2003). The ten-million word POS tagged balanced Sinica corpus of the CKIP Group was used as the training data. Several million features were expanded from the ten features mentioned above, and the training time was about two days on Windows Server 2003 x64 SP2 with an Intel Xeon 3.2GHz processor (Quad-core), 8G DRAM.

3. Results

We used the aforementioned method to perform the Taiwanese POS tagging task; nevertheless, as no standard answers were available to gauge the accuracy rate, we extracted partial results and checked them manually. The primary consideration of the manual checking procedure was the Chinese Word Segmentation and Tagging System of the CKIP group of Academia Sinica (CKIP, 2004). We selected fourteen literary works belonging to three different eras – the Ching Dynasty, the Japanese-ruled Period, and the Post-war Era. These literary works were in the form of prose (seven), novels (five), and dramas (two). We selected the first paragraph from each composition, or, if the length (number of syllables) of the first paragraph was less than 60, we selected the second paragraph.

$$accuracy\ rate = (1 - \frac{number\ of\ tagging\ errors}{number\ of\ total\ words})\ 100\% \tag{5}$$

The test data list is shown in the Appendix. Table 2 shows the test data selected for manual checking. The number of syllables, words, and incorrectly selected Mandarin words, as well as the POS tagging inaccuracy of each paragraph are noted.

A total of 1,038 words (1,496 syllables) were selected, and manual checking showed that 90 words had been incorrectly selected and 87 words were found to have inaccurate POS tagging, thus placing the average POS tagging accuracy rate at 91.6%. It should be noted that sometimes, even when the corresponding Mandarin word selected was inappropriate, the POS tagging result was still accurate. On the other hand, an appropriate or correct corresponding Mandarin word did not always have accurate POS tagging.

Furthermore, sometimes one Taiwanese word would correspond to two Mandarin words. For instance, while the Taiwanese word “壁頂/piah-téng” ‘on the wall’ is treated as only one word, the Mandarin translation “牆壁 上” should be treated as two words. There are also occasions wherein two Taiwanese words would correspond to only one Mandarin word counterpart. For instance, the Mandarin counterpart of the Taiwanese words “Tiong-kok/中國” ‘Chinese’ and “jī/字” ‘character’ was “中國字.” The former is processed as an unknown word, whereas the latter, which was separated into two independent words, was processed as two words. In these types of cases, if the POS tagging was accurate, we still regarded the results as accurate. If they were to be regarded as incorrect, the average accuracy rate would drop by around 2%.

Table 2. Tagging Accuracy Rate of The Test Data

id	No. of Syllables	No. of Words	Errors	Tagging errors	Accuracy rate(%)
1	162	109	9	6	94.5
2	66	46	4	3	93.5
3	180	119	6	8	93.3
4	122	88	3	6	93.2
5	74	51	4	1	98.0
6	75	49	7	7	85.7
7	112	87	13	12	86.2
8	101	77	7	9	88.3
9	133	93	7	9	90.3
10	116	82	3	3	96.3
11	94	59	7	5	91.5
12	92	61	10	10	83.6
13	77	59	8	4	93.2
14	92	58	3	4	93.1
Totally	1,496	1,038	90	87	91.6

In Table 3 (following), examples of actual POS tagging results are shown. It is a part of id 11. In this table, the first field is the HR mixed script and POJ script (contained in brackets), and the second field is the Mandarin candidate word(s). The "@" symbol preceding the word indicates that no entry has been found for the Taiwanese word shown in the first field; hence the HR mixed script automatically served as the Mandarin candidate word. The third field contains the selected Mandarin word, and the final field contains the selected POS. All of the incorrectly selected Mandarin words or incorrectly selected POS tags are underlined and segregated by two asterisks "**" preceding the word. The correct POS tag, contained in parentheses and shown in bold type, is then added after the incorrect POS tag.

Table 3. Example of POS Tagging Result

Taiwanese	Mandarin Candidate Words	Selected Word	POS tagging
我[góa] ‘I’	我	我	Nh
將[chiong] ‘let’	將	將	D
草帽仔[chháu-bō-á] ‘straw hat’	@草帽仔	草帽仔	Na
掛[kòa] ‘hang’	帶;掛;戴	**帶 掛	VC
tī [tī] ‘at’	在	在	P
壁頂[piah-téng] ‘on the wall’	牆壁上	牆壁上	Nc
，[,]	，	，	CommaCategory
行李[hêng-lí] ‘baggage’	行李	行李	Na
khêng[khêng] ‘arrange’	收拾;盤點	收拾	VC
khêng[khêng] ‘arrange’	收拾;盤點	收拾	VC
leh[leh] ‘(modal particle)’	咧	咧	T
，[,]	，	，	CommaCategory
坐[chē] ‘sit’	坐	坐	VA
tòa[tòa] ‘at’	住	**住	**VCL(P)
小店[sió-tiàm] ‘store’	@小店	小店	Na
ê[ê] ‘of’	的	的	DE
tha-thá-mì[tha-tha-mì] ‘tatami’	塌塌米	塌塌米	Na
頂 kôan[téng-kôan] ‘above’	上面	上面	Ncd
，[,]	，	，	CommaCategory
看[khòaⁿ] ‘see’	看	看	VC
窗外[thang-gōa] ‘outside the window’	@窗外	窗外	Nc

ê[ê] ‘of’	的	的	DE
光景[kong-kéng] ‘scenery’	風光;氣象;光景;風景;景氣	**景氣 風景	Na
，[,]	，	，	CommaCategory
看[khòaⁿ] ‘see’	看	看	VC
起起[khí-khí] ‘up’	@起起	起起	**Nb(VA)
落落[lo̍h-lo̍h] ‘down’	@落落	落落	VA
ê[ê] ‘of’	的	的	DE
海湧[hái-éng] ‘wave’	海浪 ;海潮	海浪	Na
，[,]	，	，	CommaCategory
因爲[in-ūi] ‘because’	由於 ;因爲	因爲	Cbb
等待[tán-thāi] ‘wait’	留待 ;等待	等待	VK
朋友[pêng-iú] ‘friend’	友人 ;朋友	朋友	Na
，[,]	，	，	CommaCategory
心適[sim-sek] ‘pleasant’	好玩 ;好玩兒 ;有趣 ;風趣 ;愉快 ;稀奇 ;鬧著玩	有趣	VH
心適[sim-sek] ‘pleasant’	好玩 ;好玩兒 ;有趣 ;風趣 ;愉快 ;稀奇 ;鬧著玩	有趣	VH
，[,]	，	，	CommaCategory
輕輕仔[khin-khin-á] ‘lightly’	輕輕的	輕輕的	**Nb(D)
來[lâi] ‘toward’	來	來	D
點[tiám] ‘light’	燃點;檢點;點;點子	點	VC
一支[chi̍t-ki] ‘a’	@一支	一支	Na
涼涼[liâng-liâng] ‘cool’	冷冷;涼絲絲	**冷冷 涼涼	VH
ê[ê] ‘of’	的	的	DE
芎蕉[kin-chio] ‘banana’	香蕉	香蕉	Na
薰[hun] ‘tobacco’	香菸;香煙;薰	香煙	Na
。[.]	。	。	PeriodCategory

4. Error Analysis

This section discusses how a more thorough check was performed to analyze the error conditions.

4.1 Selection of Inappropriate Mandarin Word

An analysis of the errors made in the selection of Mandarin words or POS tags revealed that the selection of inappropriate Mandarin words led to POS tagging errors in 25 cases. Table 4 shows the incorrect Mandarin words selected and their respective POS.

Table 4. The Selected Incorrect Mandarin Words and Their Respective POS

Word	Selected Mandarin word and POS	More appropriate Mandarin word and POS	Remark
押/ah	強制(D) ‘compel’	押(VC) ‘take into custody’	
bat/bat	知道(VK) ‘know’	曾(D) ‘ever’	
無/bô	不(D) ‘not’	沒有(VJ) ‘not have’	2 times
chham/chham	和(P) ‘and’	摻(VC) ‘accompany’	
進前 /chìn-chêng	之前(Ng) ‘before’	向 前(P Nes) ‘forward’	
這號/chit-hō	這樣(VH) ‘such’	這種(Nep Nf) ‘this kind of’	2 times
轉/chōan	上(Ncd) ‘above’	轉(Vac) ‘turn’	2 times
外/gōa	外(Ng) ‘outside’	開外(Neqa) ‘more’	2 times
夭壽/iáu-siū	非常(Dfa) ‘very’	早夭(VH) ‘dead early’	
加/ke	上(Ncd) ‘above’	多(Dfa) ‘more’	
價值/kè-tȧt	值得(VH) ‘worthy’	價值(Na) ‘value’	
腳/kha	個(DE) ‘(a numerary adjunct)’	下(Ncd) ‘under’	
黃 hóaⁿ/n̂g-hóaⁿ	罕(D) ‘rarely’	淺黃(A) ‘light yellow’	
倚/óa	依(P) ‘in accordance with’	靠(VJ) ‘lean against’	
活/ȯah	生活(Na) ‘life’	活(VH) ‘live’	
破相/phòa-siùⁿ	破(VHC) ‘break’	殘廢(Na) ‘disabled’	
細漢/sè-hàn	小時候(Nd) ‘in one's childhood’	年幼(VH) ‘young’	
相借問 /sio-chioh-mn̄g	招呼(VC) ‘greet’	打招呼(VB) ‘say hello’	
搭/tah	地方(Na) ‘location’	搭(VC) ‘construct’	
tiȯh/tiȯh	就(P) ‘(an auxiliary confirming and stressing the verb following)’	著(VCL) ‘come into contact with’	
著/tiȯh	就(P) ‘(an auxiliary confirming and stressing the verb following)’	得(D) ‘need to’	

4.2 Absence of Appropriate Mandarin Translation in OTMD

There were fourteen errors made in inappropriate Mandarin word selection due to the absence of an appropriate Mandarin word in the OTMD. This also led to errors in the POS tagging. The discovery indicates the necessity of expanding the entries of the OTMD. Table 5 tabulates these errors.

Table 5. Errors Caused by Absence of Appropriate Mandarin Word Option in OTMD

Taiwanese	Selected Mandarin by System	Appropriate Mandarin Word	Remark
chak/chak	促(VF) ‘urge’	擠(VC) ‘crowd’	
chūn/chūn	絞(VC) ‘twist’	陣(Nf) ‘(a numerary adjunct)’	2 times
kah/kah	和(Caa) ‘and’	得(DE) ‘a particle used after a verb’	3 times
leh/leh	咧(T) ‘(modal particle)’	在(P) ‘doing’	3 times
煞/soah	結束(VHC) ‘finish’	卻(D) ‘but’	
teh/teh	在(P) ‘(an indicator or location)’	著(Di) ‘(an adverbial particle)’	
頂/téng	頂(VC) ‘lift’	上(Nes) ‘(the first half part)’	
tiāⁿ-tiāⁿ / tiāⁿ-tiāⁿ	常常(D) ‘often’	而已(T) ‘just’	
轉 / tńg	調解(VC) ‘mediate’	轉(VAC) ‘turn’	

4.3 Unknown Words from the Viewpoint of Mandarin

Ten of the POS tagging errors were made because the word was an unknown word. Parts of these unknown words correspond to two Mandarin words. These unknown words are tabulated in Table 6.

Table 6. Unknown Words from The Viewpoint of Mandarin

Taiwanese Word	Corresponding Mandarin Word	Selected POS by System	Correct POS
bē 會/bē-ē	不會 ‘be unable to’	Nb	D
廟埕/biō-tiâⁿ	廟前院 ‘temple square’	Na	Nc (Na Nc)
食老/chia̍h-lāu	年老 ‘old’	Na	VH
轉了/chōan-liáu	轉 後 ‘after turning’	VH	VC Ng
牛擔灣/Gû-taⁿ-oan	牛擔灣 ‘(a place name)’	VA	Nc
法律上/hoat-lu̍t-siōng	法律上 ‘jural’	VC	N (Na Ncd)
非爲/hui-ûi	非爲 ‘infamous conduct’	A	N (A Na)
窮志/kiông-chì	窮志 ‘exhaust the ambition’	Na	V (VH Na)
輕輕仔/khin-khin-á	輕輕地 ‘lightly’	Nb	D (VH DE)
生子/seⁿ-kiáⁿ	生孩子 ‘give birth to a child’	Na	VA (VH Na)

4.4 Propagation Error

Five of the POS tagging errors were probably due to the occurrence of a previous POS tagging error. These are categorized as propagation errors and include one unknown word.

4.5 Other Cases

The personal name “天賜” of “天賜 ah/Thian-sù ah” (not an unknown word) which has been tagged as “A” with the suffix “ah” tagged as “T” or “Di” (which appeared twice in all; once, the selected Mandarin word was “啊” and in other instance it was “了”).

The Taiwanese word “對/tùi” under general circumstances is synonymous with the Mandarin word “從”‘from.’ This word appeared ten times in the test data. The system selected the Mandarin word “對”‘for’ eight times and the word “從” twice for its counterpart. Nevertheless, under both circumstances, the POS tag of the word was always “P”; thus, the different word choice did not affect the accuracy of the POS tagging.

There were also 30 errors made that leave us unable to clearly explain the reasons. Table 7 lists some examples.

Table 7. Example of Some POS Tagging Errors

Left Context	Word and POS	Correct POS	Right Context	id
	lūn ‘discuss’ (Na)	VE	tha̍k‘read’(VC) pe̍h-ōe-jī ‘vernacular writing’(Na) khah-iâⁿ‘better than’(VJ) …	1
chò ‘do’(VC) thâi-lâng ‘kill someone’(VA)	hōan ‘criminal’ (VC)	Na	siū ‘be subjected to’(P) sí-hêng ‘death penalty’(Na) ê‘of’(DE) 7(Neu) lâng ‘people’(Na)	2
lâng‘people’(Na)	chi̍t-ē ‘once’ (Nd)	D	chia̍h-lāu ‘old’(VH)	3
tùi ‘from’ (P)	khí-thâu ‘beginning’(VH)	Nv	chiū‘then’(D) chin‘very’(Dfa) tāng ‘waver’(VAC)	4
Má-lī ‘a person name’(Nb) ê‘of’(DE) lāu-pē‘father’(Na)	sí ‘dead’ (Dfb)	VH	ê‘of’(DE) sî‘time’(Na)	10
khòaⁿ ‘look at’ (VC)	khí-khí‘up’(Nb)	VA	lo̍h-lo̍h‘down’(VA) ê‘of’(DE) hái-éng ‘tide’(Na)	11
n̂g-hóaⁿ ‘turned yellow’(VH) àm-tām‘dim’(VH) ê‘of’(DE) lō͘-teng‘streetlamp’(Na)	chhiō ‘shine’(D)	VC	lóng‘always’(D) bē‘not’(D) hn̄g‘far’(VH)	12

4.6 Summary of Error Conditions

A summary of the causes of the errors made during the POS tagging and their frequency percentages is tabulated in Table 8.

Table 8. The Reason of POS Tagging Errors

Reason	Count	Percentage(%)	Remark
Selection of Inappropriate Mandarin Word	25	28.7	
Absence of Appropriate Mandarin Word	14	16.1	
Unknown Word	10	11.5	
Personal Name	4	4.6	
Propagation Error	4	4.6	Includes an unknown word
Totally	57	65.5	After discounting the repeat count

5. Discussion

5.1 Is Improvement Possible?

The ideal situation would be to resolve the foregoing errors and use this method to conduct the Taiwanese POS tagging to achieve an accuracy rate of 97.1%. Nevertheless, there is an apparent difficulty in the realization of this goal.

There are differences between the Taiwanese word order and the Mandarin word order; thus, the selection of the incorrect Mandarin word, and consequently incorrect POS tagging, occurred with high probability. The absence of appropriate Mandarin translation was the second leading cause of the POS tagging errors.

The unknown word problem was also a cause of POS tagging errors. From the Mandarin perspective, these words are not actually unknown words; this problem mostly resulted from the fact that translations between different languages are not one-to-one mappings. Another significant factor involves the use of hyphens in the POJ script, as their usage has not yet been standardized. It is probable that due to the use of Han characters, word boundaries are relatively vague in the different languages of the Chinese language family.

5.2 Hyphen Problems, Distinction between Taiwanese and Mandarin

In Taiwanese, some words take on the POJ script, thus, the use of the hyphen. Used one way, they separate the syllables of words, making it possible for a syllable to correspond to a Han character; used another way, they serve as word separators. Each syllable in a hyphenated word represents a unigram, and a space separates each word. Unfortunately, no original word boundaries of Han character writing can be found to correspond to the hyphenated word.

In addition, Taiwanese has around 3,000 legal syllables, whereas Mandarin has around 1,200 legal syllables (Chan, 2008). Because of this, it may be said that the Taiwanese language has more single-syllable words. Nevertheless, as a single-syllable word may have several corresponding Han characters, the use of two-syllable or multi-syllable words resolves most of the problems.

For instance, if the Taiwanese word "這個"'this one' is written as "chit ê" (no hyphen used), the syllable "chit" may be made to correspond to several Mandarin words, such as "這" 'this,' "職"'job,' "質" 'quality,' "織," 'knit,' *etc*. The syllable "ê" may also be made to correspond to several Mandarin words, such as "的" 'of,' "個" '(a numerary adjunct),' "鞋" 'shoe,' *etc*. If the word is written as "chit-ê" (hyphenated), it definitely corresponds to "這個" in HR script. Hence, under the POJ script, the writer may tend to use a hyphen to link a single-syllable word to another single-syllable word if these two single-syllable words may likely form one composite word or one phrase. Present practices show that the word "這個" may appear hyphenated or in a separated syllable form, thus creating inconsistencies.

As the use of hyphenated words creates the problem of one Taiwanese word corresponding to two Mandarin words, if the original text is not revised and the Mandarin corresponding word is manifested as an unknown word, it may be possible to just remove the hyphen and try again. This method may reduce the chance of POS tagging errors due to the unknown word factor.

5.3 The Distinction between Different Eras or Different Genres

We investigated whether texts of a different era or a different literary genre would affect the accuracy rate of the POS tagging. Table 10 shows the POS tagging accuracy rates for texts of three types of literary genres and Table 11 shows the POS tagging accuracy rates for texts of literary works belonging to three different periods or eras. Table 9 shows that the POS tagging accuracy rate for novel materials is comparably lower than other genres; whereas Table 10 indicates that the POS tagging accuracy rate for the materials written in the Post-war era are comparably lower than the other periods investigated. Basically, there are no significant differences among three genres or three eras as a whole.

Table 9. Tagging Accuracy Rate for Different Genres

Genre	No. of Words	No. of Tagging Errors	Accuracy Rate (%)
Prose	549	43	92.2
Novel	372	36	90.3
Drama	117	8	93.2

Table 10. Tagging Accuracy Rate for Different Eras

Era	No. of Words	No. of Tagging Errors	Accuracy Rate (%)
Ching Dynasty	232	18	92.2
Japanese-ruled	359	27	92.5
Post-war	447	42	90.6

After deliberation, we found that the individual writing style of authors is actually the dominant factor of the POS tagging accuracy. From Table 2, the individual POS tagging accuracy varies from 83.6% to 98.0%.

6. Conclusion and Future Works

We proposed a Taiwanese POS tagging method using a statistical method and Mandarin training data, and we achieved an accuracy rate of 91.6%. Due to the lack of Taiwanese training data, we sought the help of Mandarin.

This strategy could also be applied to other languages that lack resources. We think that this is a very important idea. It is preferable to select an intermediate language close to the target language from the viewpoint of the language family.

We also developed an online Taiwanese word segmentation and POS tagging system for people who are interested in this topic. Users can input Taiwanese text and get the POS tagging results. It is somewhat difficult for a user to prepare both POJ and HR mixed scripts; therefore, we also provide the functions in the absence of one of these two scripts (Iunn, *et. al.*, 2007). This, however, will decrease the accuracy rate.

If we can construct a Taiwanese-Mandarin parallel corpus, we can use other methods like the Coerced Markov Models proposed by Fung and Wu (1995) to accomplish the Taiwanese POS tagging task.

We hope that we can proceed to the construction of Taiwanese Treebank.

Acknowledgments

This research was supported in part by National Science Council of Taiwan, under the contract number NSC 95-2221-E-122 -006. Thanks to the National Museum of Taiwanese Literature of Taiwan for providing plentiful written Taiwanese literature data. We also thank the anonymous reviewers for their constructive opinions.

Reference

Berger, A. L., Pietra, S. A. D., & Pietra, V. J. D. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1), 39-71.

Chan, K. I. (2008). *Comparison with the Usage of Academic and Non-academic Taiwanese Words*. Master thesis, National Taitung University.

Chou, S. Y. (2006). *T3 Taiwanese Treebank and Brill Part-of-Speech Tagger*. Master thesis, Hsin-chu: National Tsing Hua University.

CKIP. (1993). *Analysis of Chinese Part-of-speech.* The Association for Computational Linguistics and Chinese Language.

CKIP. (2004). *Chinese Word Segmentation and Tagging System*, (Retrieved 2009/4/10) http://ckipsvr.iis.sinica.edu.tw/.

Embree, B. L. M. (1984). *A dictionary of Southern Min* '台英辭典,' Taipei: Taipei Language Institute.

Fung, P., & Wu, D. K. (1995). Coerced Markov Models for cross-lingual lexical tag relations. in the *6th International Conference on Theoretical and Methodological Issues in Machine Translation*, 1, 1995, Leuven, Belgium, 240-255.

Gordon, R. G. Jr. ed. (2005). *Ethnologue: Languages of the world* (15th ed.). Dallas:SIL International.

Huang, S. F.(1995). *Language, Society and Ethnicity* (2nd ed.). Taipei: Crane.

Iunn, U. G. (2000). *Online Taiwanese-Mandarin Dictionary*. (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/q/q.asp.

Iunn, U. G. (2003a). *Online Taiwanese Syllable Dictionary.* (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/TG/jitian/.

Iunn, U. G. (2003b). Survey of the Online Taiwanese-Mandarin Dictionary-- Discussion of Building Technique and its Utilization. in the *Proceedings of 3rd International Conference on Internet Chinese Education*, 2003b, Overseas Chinese Affairs Commission, 132-141.

Iunn, U. G. (2003c). *Online Taiwanese Concordancer System.* (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/TG/concordance/.

Iunn, U. G. (2005). *Taiwanese Corpus Collection and Corpus Based Syllable / Word Frequency Counts for Written Taiwanese*. (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp.

Iunn, U. G. (2007). New Manifestation of the Taiwanese vernacular literature -- Introduction to Digital Archive for Written Taiwanese. *National Museum of Taiwanese Literature Communication*, 15, 42-44.

Iunn, U. G. (2009). *Processing Techniques for Written Taiwanese-- Tone Sandhi and POS Tagging*. PhD thesis, National Taiwan University.

Iunn, U. G., & Lau, K. G. (2007). Introduction to online Taiwanese Dictionaries and Corpora. *Language, Society and Culture Series 2: Multiculturalism Thinking of the Language Policy*, 2007, Institute of Linguistics of Academia Sinica, 311-328.

Iunn, U. G., Lau, K. G., Tan-Tenn, H. G., Lee, S. A., & Kao, C. Y. (2007). Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods. *International*

Journal of Computational Linguistics and Chinese Language Processing, 12(4), 349-370.

Iunn, U. G., Lau, K. G., & Tai, C. H. (2007). *Online Taiwanese Word Segmentation and POS Tagging System*, (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/TGB/tagging/tagging.asp.

Lau, K. G. (2007). *Finding Mandarin Candidate Words by POJ script and Han-Romanization mixed script word pair*, (Retrieved 2009/4/10) http://iug.csie.dahan.edu.tw/nmtl/dadwt/pos_tagging/clhl_hoagi_hausoansu.asp.

Le, Z. (2003). Maximum Entropy Modeling Toolkit for Python and C++. (Retrieved 2009/7/10) http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge: MIT Press.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. in the *Proceedings of 17th International Conference on Machine Learning*, Stanford University, 591-598.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. in the *Proceedings of the IEEE*, 77, 257-286.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania, 133-142.

Samuelsson, C. (2003). Statistical methods. in the *Oxford Handbook of Computational Linguistics*, Oxford University Press, 358-375.

Shi, D. M. (2006). *T3 Taiwanese Treebank and Brill Parser*, Master thesis, Hsin-chu: National Tsing Hua University.

Tai, C. H. (2007). *Word and POS tagging selection for Taiwanese Language*, (Retrieved 2009/4/10) http://140.109.19.105/.

Tsai, Y. F., & Chen, K. J. (2004). Reliable and Cost-Effective Pos-Tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1), 2004, 83-96.

Appendix

Test Data List.

id	Year	Genre	Author	Article title	No. of Syllables
1	1885	prose	Reverend Ia̍p ‘葉牧師’	Pe̍h-ōe-jī ê lī-ek ‘The Benefits of Using Pe̍h-ōe-jī，白話字的利益’	162
2	1893	prose	Reverend Kam ‘甘牧師’	Chhiⁿ-mî o̍h ‘Blind Study，青瞑學’	66
3	1919	prose	H S K	Phín-hēng ê ûi-thôan ‘Inheritance of Morality，品行的遺傳’	180
4	1935	prose	Ong Chong-têng ‘汪宗程’	Chín-chai kì ‘Earthquake Disaster Record，震災記’	122
5	1954	prose	Ô͘ Bûn-tî ‘胡文池’	Tōa-soaⁿ chhiùⁿ-koa ‘A High Mountains sing，大山唱歌’	74
6	1990	prose	Tân Gī-jîn ‘陳義仁’	Lāu-lâng ê kè-ta̍t ‘The Value of The Elderly People，老人的價值’	75
7	2000	prose	Tân Bêng-jîn ‘陳明仁’	Sûn-chêng Ông Pó-chhoan ‘Pure Love Ông Pó-chhoan，純情王寶釧’	112
8	1890	novel	Unknown	An-lo̍k-ke ‘Safety and Happiness Street，安樂街’	101
9	1924	novel	Lōa Jîn-seng ‘賴仁聲’	Án-niá ê Ba̍k-sái ‘Mother's Tears，母親的眼淚’	133
10	1955	novel	N̂g Hôai-un ‘黃懷恩’	Chháu-tui téng ê bîn-bāng ‘Dreams on the Grass Stack，草堆上的夢’	116
11	1990	novel	Iûⁿ Ún-giân ‘楊允言’ translated	Hái-phīⁿ Sin-niû ‘Bride on The Cape，岬角上的新娘’	94
12	2006	novel	Lâu Sêng-hiân ‘劉承賢’	Chia̍h-chōe ‘Plead Guilty，伏罪’	92
13	1924	drama	Lîm Bō͘-seng ‘林茂生’	Hì-chhut: Lō͘-tek kái kàu ‘ Drama: Ruth Reformed Church，戲齣:路得改教’	77
14	1950	drama	Tân Chheng-tiong ‘陳清忠’ translated	Venice ê Seng-lí-lâng ‘Venice Businessman，威尼斯的生意人’	92
Note: the original author of id 11 is Sòng Te̍k-lâi ‘宋澤萊,’ id 14 is Shakespeare					

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 3, September 2009, pp. 257-280

A Thesaurus-Based Semantic Classification of English Collocations

Chung-Chi Huang*, Kate H. Kao+, Chiung-Hui Tseng+ and Jason S. Chang+

Abstract

Researchers have developed many computational tools aimed at extracting collocations for both second language learners and lexicographers. Unfortunately, the tremendously large number of collocates returned by these tools usually overwhelms language learners. In this paper, we introduce a thesaurus-based semantic classification model that automatically learns semantic relations for classifying adjective-noun (A-N) and verb-noun (V-N) collocations into different thesaurus categories. Our model is based on iterative random walking over a weighted graph derived from an integrated knowledge source of word senses in *WordNet* and semantic categories of a thesaurus for collocation classification. We conduct an experiment on a set of collocations whose collocates involve varying levels of abstractness in the collocation usage box of Macmillan English Dictionary. Experimental evaluation with a collection of 150 multiple-choice questions commonly used as a similarity benchmark in the TOEFL synonym test shows that a thesaurus structure is successfully imposed to help enhance collocation production for L2 learners. As a result, our methodology may improve the effectiveness of state-of-the-art collocation reference tools concerning the aspects of language understanding and learning, as well as lexicography.

Keywords: Collocations, Semantic Classification, Semantic Relations, Random Walk Algorithm, Meaning Access Index and *WordNet*.

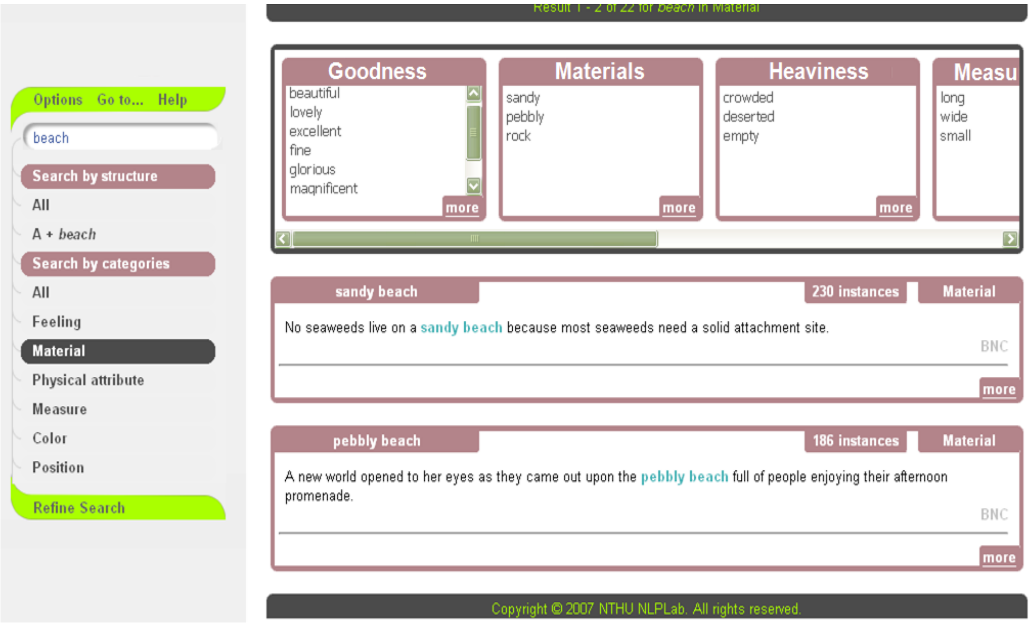
* CLCLP, TIGP, Academia Sinica, Taipei, Taiwan
+ Institute of Information Systems and Applications, NTHU, Hsinchu, Taiwan
E-mail: {u901571, msgkate, smilet, jason.jschang}@gmail.com

1. Introduction

Researchers have developed applications of computational collocation reference tools, such as several commercial collocation dictionary CD-ROMs, Word Sketch (Kilgarriff & Tugwell, 2001), *TANGO* (Jian *et al*., 2004), to answer queries (*e.g*., a search keyword “beach” for its adjective collocates) of collocation usage. These reference tools typically return collocates (*e.g*., adjective collocates for the pivot word “beach” are “rocky,” “golden,” “beautiful,” “raised,” “sandy,” “lovely,” “unspoiled,” “magnificent,” “deserted,” “fine,” “pebbly,” “splendid,” “crowded,” “superb,” *etc*.) extracted from a corpus of English texts (*e.g*., *British National Corpus*).

Unfortunately, existing tools for language learning sometimes present too much information in a batch on a single screen. With corpus sizes rapidly growing to Web scale (*e.g*., Web 1 Trillion 5-gram Corpus), it is common to find hundreds of collocates for a query word. The bulk of information may frustrate and slow L2 learners’ progress of learning collocations. An effective language learning tool also needs to take into consideration second language learners’ absorbing capacity at one sitting. To satisfy the need for presenting a digestible amount of information at one time, a promising approach is to automatically partition collocations of a query word into various categories to support meaningful access to the search results and to give a thesaurus index to collocation reference tools.

Consider the query “beach” in a search for its adjective collocates. Instead of generating a long list of adjectives like the above-mentioned applications, a better presentation could be composed of clusters of adjectives inserted into distinct semantic categories such as: {*fine, lovely, superb, beautiful, splendid*} assigned with a semantic label “*Goodness,*” {*sandy, rocky, pebbly*} assigned with a semantic label “*Materials,*” *etc*. Intuitively, by imposing a semantic structure on the collocations, we can bias the existing collocation reference tools towards giving a thesaurus-based semantic classification as one of the well-developed and convincingly useful collocation thesauri. We present a thesaurus-based classification system that automatically groups collocates of a given pivot word (here, the adjective collocates of a noun, the verb collocates of a noun, and the noun collocates of a verb) into semantically related classes expected to render highly useful applications in computational lexicography and second language teaching for L2 learners. A sample presentation for a collocation thesaurus is shown in Figure 1.



No seaweeds live on a sandy beach because most seaweeds need a solid attachment site.

A new world opened to her eyes as they came out upon the pebbly beach full of people enjoying their afternoon promenade.

Figure 1. Sample presentation for the adjective collocate search query "beach".

Our thesaurus-based semantic classification model has determined the best semantic labels for 859 collocation pairs, focusing on: (1) A-N pairs and clustering over the adjectives (*e.g.*, "fine beach"); (2) V-N pairs and clustering over the verbs (*e.g.*, "develop relationship"); and (3) V-N pairs and clustering over the nouns (*e.g.*, "fight disease") from the specific underlying collocation reference tools (in this study, from *JustTheWord*). Our model automatically learns these useful semantic labels using the Random Walk Algorithm, an iterative graphical approach, and partitions collocates for each collocation types (*e.g.*, the semantic category "*Goodness*" is a good thesaurus label for "fine" in the context of "beach" along with other adjective collocates such as "lovely," "beautiful," "splendid," and "superb"). We describe the learning process of our thesaurus-based semantic classification model in more detail in Section 3. At runtime, we assign the most probable semantic categories to collocations (*e.g.*, "sandy," "fine," "beautiful," *etc.*) of a pivot word (*e.g.*, "beach") for semantic classification. In this paper, we exploit the Random Walk Algorithm to disambiguate word senses, assign semantic labels, and partition collocates into meaningful groups.

The rest of the paper is organized as follows. We review the related work in the next section. Then, we present our method for automatic learning to classify collocations into semantically related categories, which is expected to improve the presentation of underlying collocation reference tools and support collocation acquisition by computer-assisted language learning applications for L2 learners (Section 3). As part of our evaluation, two metrics are designed with very little precedent of this kind. One, we assess the performance of resulting collocation clusters by a robust evaluation metric; two, we evaluate the conformity of semantic labels by a three-point rubric test over a set of collocation pairs chosen randomly from the classifying results (Section 5).

2. Related Work

Many natural language processing (NLP) applications in computational lexicography and second language teaching (SLT) build on one part of lexical acquisition emphasizing teaching collocation for L2 learners. In our work, we address an aspect of word similarity in the context of a given word (*i.e.*, collocate similarity), in terms of use, acquisition, and ultimate success in language learning.

This section offers the theoretical basis on which recommendations for improvements to the existing collocation reference tools are made, and it is made up of three major sections. In the first section, an argument is made in favor of collocation ability being an important part of language acquisition. Next, we show the need to change the current presentation of collocation reference tools. The final section examines other literature on computational measures for word similarity versus collocate similarity.

2.1 Collocations for L2 Learners

The past decade has seen an increasing interest in the studies on collocations. This has been evident not only from a collection of papers introducing different definitions of the term “collocation” (Firth, 1957; Benson, 1985; Nattinger & DeCarrico, 1992; Nation, 2001), but also from the inclusive review of research on collocation teaching and the relation between collocation acquisition and language learning (Lewis, 1997; Hall, 1994).

New NLP applications for extracting collocations, therefore, are a great boon to both L2 learners and lexicographers alike. SLT has long favored grammar and memorization of lexical items over learning larger linguistic units (Lewis, 2000). Nevertheless, several studies have shown the importance of acquisition of collocations; moreover, they have found specifically that the most important is learning the right verbs in verb-noun collocations (Nesselhauf, 2003; Liu, 2002). Chen (2004) showed that verb-noun (V-N) and adjective-noun (A-N) collocations were found to be the most frequent error patterns. Liu (2002) found that, in a study of English learners’ essays from Taiwan, 87% of miscollocations were attributed to the misuse of V-N collocations. Of those, 96% were due to the selection of the wrong verb. A simple example will suffice to illustrate: in English, one writes a check and also writes a letter while the equivalent Mandarin Chinese word for the verb “write” is “kai” (開) for a check and “xie” (寫) for a letter, but absolutely not “kai” (開) for a letter.

This type of language-specific idiosyncrasy is not encoded in either pedagogical grammars or lexical knowledge but is of utmost importance to fluent production of a language.

2.2 Meaning Access Indexing in Dictionaries

Some attention has been paid to the investigation of the dictionary needs and reference skills of language learners (Scholfield, 1982; Béjoint, 1994), and one important cited feature is a structure to support users' neurological processes in meaning access. Tono (1984) was among the first attempts to claim that the dictionary layout should be more user-friendly to help L2 learners access desired information more effectively. According to Tono (1992) in his subsequent empirical close examination of the matter, menus that summarize or subdivide definitions into groups at the beginning of entries in dictionaries would help users with limited reference skills to access the information in the dictionary entries more easily. The *Longman Dictionary of Contemporary English*, 3rd edition [ISBN 0-582-43397-5] (henceforth called *LDOCE3*), has just such a system called "**Signposts**". When words have various distinct meanings, the *LDOCE3* begins each sense anew with a word or short phrase which helps users more effectively discover the meaning they need. The *Cambridge International Dictionary of English* [ISBN 0-521-77575-2] does this as well, creating an index called "**Guide Word**" which provides similar functionality. Finally, the *Macmillan English Dictionary for Advanced Learners* [ISBN 0-333-95786-5], which has "Menus" for heavy-duty words with many senses, utilizes this approach as well.

Therefore, in this paper, we introduce a classification model for imposing a thesaurus structure on collocations returned by existing collocation reference tools, aiming at facilitating concept-grasping of collocations for L2 learners.

2.3 Similarity of Semantic Relations

The construction of practical, general word sense classification has been acknowledged to be one of the most difficult tasks in NLP (Nirenburg & Raskin, 1987), even with a wide range of lexical-semantic resources such as *WordNet* (Fellbaum, 1998) and *Word Sketch* (Kilgarriff & Tugwell, 2001).

Lin (1997) presented an algorithm for word similarity measured by its distributional similarity. Unlike most corpus-based word sense disambiguation (WSD) algorithms, where different classifiers are trained for separate words, Lin used the same local context database as the knowledge source for measuring all word similarities. Approaches presented to recognize synonyms have been studied extensively (Landauer & Dumais, 1997; Deerwester *et al.*, 1990; Turney, 2002; Rehder *et al.*, 1998; Morris & Hirst, 1991; Lesk, 1986). Measures of recognizing collocate similarity, however, are not as well developed as measures of word similarity.

The most closely related work focuses on automatically classifying semantic relations in noun pairs (*e.g.*, mason:stone) and evaluation with a collection of multiple-choice word analogy question from the SAT exam (Turney, 2006). Another related approach, presented in Nastase and Szpakowicz (2003), describes how to automatically classify a noun-modifier pair, such as “laser printer,” according to the semantic relation between the head noun (printer) and the modifier (laser). The evaluation is manually conducted by human labeling. For a review of work to a more fine-grained word classification, Pantel and Chklovski (2004) presented a semi-automatic method for extracting fine-grained semantic relations between verbs. VerbOcean (http://semantics.isi.edu/ocean/) is a broad-coverage semantic network of verbs, detecting similarity (*e.g.*, transform::integrate), strength (*e.g.*, wound::kill), antonymy (*e.g.*, open::close), enablement (*e.g.*, fight::win), and temporal happens-before (*e.g.*, marry::divorce) relations between pairs of strongly associated verbs using lexico-syntactic pattern over the Web. Hatzivassiloglou and McKeown (1993) presented a method towards the automatic identification of adjectival scales. Based on statistical techniques with linguistic information derived from the corpus, the adjectives, according to their meaning based on a given text corpus, can be placed in one group describing different values of the same property. Their clustering algorithm suggests some degree of adjective scalability; nevertheless, it is interesting to note that the algorithm discourages recognizing the relationship among adjectives, *e.g.*, missing the semantic associations (for example a semantic label of “time associated”) between *new-old*. More recently, Wanner *et al*. (2006) sought to semi-automatically classify the collocations from corpora via the lexical functions in dictionary as the semantic typology of collocation elements. While there is still a lack of fine-grained semantically-oriented organization for collocation, *WordNet* synset (*i.e.*, synonymous words in a set) information can be explored to build a classification scheme for refinement of the model and develop a classifier to measure the distribution of class for the new tokens of words set foot in. Our method, which we will describe in the next section, uses a similar lexicon-based approach for a different setting of collocation classification.

3. Methodology

3.1 Problem Statement

We focus on the preparation step of partitioning collocations into categories for collocation reference tools: providing words with semantic labels, thus, presenting collocates under thesaurus categories for ease of comprehension. The categorized collocations are then returned in groups as the output of the collocation reference tool. It is crucial that the collocation categories be fairly consistent with human judgment and that the categories of collocates cannot be so coarse-grained that they overwhelm learners or defeat the purpose of users’ fast access. Therefore, our goal is to provide semantic-based access to a well-founded collocation thesaurus. The problem is now formally defined.

Problem Statement: We are given (1) a set of collocates *Col* = {C_1, C_2, …, C_n} (*e.g.*, “sandy,” “beautiful,” “superb,” “rocky,” *etc.*) with corresponding parts-of-speech $P=\{p|\ p \in Pos$ and $Pos=\{$noun,adjective,verb$\}\}$ for a pivot word *X* (*e.g.*, “beach”); (2) a combination of thesaurus categories (*e.g.*, *Roget’s Thesaurus*), *TC* = {(*W*, *P*, *L*)} where a word *W* with a part-of-speech *P* is under the general-purpose semantic category *L* (*e.g.*, feelings, materials, art, food, time, etc.); and (3) a lexical database (*e.g.*, *WordNet*) as our word sense inventory *SI* for semantic relation population. *SI* is equipped with a measure of semantic relatedness: REL(*S*, *S*’) encodes semantic relations holding between word sense *S* and *S*’.

Our goal is to partition *Col* into subsets of similar collocates by means of integrated semantic knowledge crafted from the mapping of *TC* and *SI*, whose elements are likely to express related meanings in the same context of *X*. For this, we leverage a graph-based algorithm to assign the most probable semantic label *L* to each collocation, thus giving collocations a thesaurus index.

For the rest of this section, we describe our solution to this problem. In the first stage of the process, we introduce an iterative graphical algorithm for providing each word with a word sense (Section 3.2.1) to establish integrated semantic knowledge. A mapping of words, senses, and semantic labels is thus constructed for later use of automatic collocation partitioning. In the second stage (Section 3.2.2), to reduce out-of-vocabulary (OOV) words in *TC*, we extend word coverage of limited *TC* by exploiting a lexical database (*e.g.*, *WordNet*) as a word sense inventory, encoding words grouped into cognitive synonym sets and interlinked by semantic relations. In the third stage, we present a similar graph-based algorithm for collocation labeling using the extended *TC* and Random Walk on a graph in order to provide a semantic access to collocation reference tools of interest (Section 3.3). The approach presented here is generalizable to allow construction from any underlying semantic resource. Figure 2 shows a comprehensive framework for our unified approach.

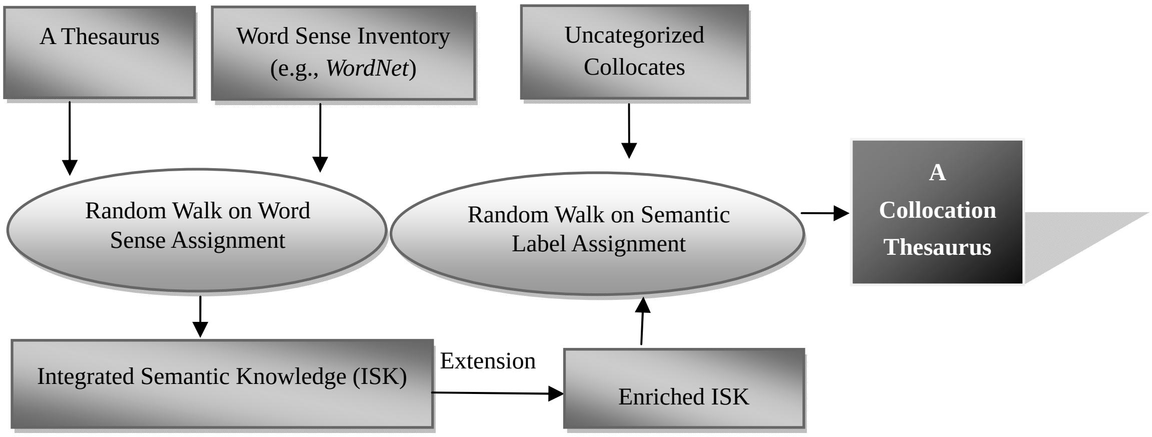

A Thesaurus
Word Sense Inventory (e.g., *WordNet*)
Uncategorized Collocates
Random Walk on Word Sense Assignment
Random Walk on Semantic Label Assignment
A Collocation Thesaurus
Integrated Semantic Knowledge (ISK)
Extension
Enriched ISK

Figure 2. A comprehensive framework for our classification model.

3.2 Learning to Build a Semantic Knowledge by Iterative Graphical Algorithms

In this paper, we attempt to provide each word with a semantic label and attempt to partition collocations into thesaurus categories. In order to partition a large-scale collocation input and reduce the out-of-vocabulary (OOV) encounters for the model, we first incorporate word sense information in *SI*, into the thesaurus, *i.e.*, *TC*, and extend the former integrated semantic knowledge (*ISK*) using semantic relations provided in *SI*. Figure 3 outlines the aforementioned process.

(1) Build an Integrated Semantic Knowledge (*ISK*) by Random Walk on Graph (Section 3.2.1)

(2) Extend Word Coverage for Limited *ISK* by Lexical-Semantic Relations (Section 3.2.2)

Figure 3. Outline of the learning process of our model.

3.2.1 Word Sense Assignment

In the first stage (Step (1) in Figure 3), we use a graph-based sense linking algorithm which automatically assigns appropriate word senses to words under a thesaurus category. Figure 4 shows the algorithm.

Algorithm 1. Graph-based Word Sense Assignment

Input: A word list, *WL*, under the same semantic label in the thesaurus *TC*; A word sense inventory *SI*.

Output: A list of linked word sense pairs, {(*W*, *S**)}

Notation: Graph *G* = {*V*, *E*} is defined over admissible word senses (*i.e.*, *V*) and their semantic relations (*i.e.*, *E*). In other words, each word sense *S* constitutes a vertex $v \in V$ while a semantic relation between senses *S* and *S'* (or vertices) constitutes an edge in *E*. Word sense inventory *SI* is organized by semantic relations *SR* and REL(*S*,*S'*) identifies the semantic relations between sense of *S* and *S'* in *SI*.

PROCEDURE AssignWordSense(*WL*,*SI*)

Build weighted graph *G* of word senses and semantic relations

```
        INITIALIZE $V$ and $E$ as two empty sets
        FOR each word $W$ in $WL$
            FOR each of the $n(W)$ admissible word senses, $S$, of $W$ in $SI$
(1)             ADD node $S$ to $V$
        FOR each node pair $(S,S')$, where $S$ and $S'$ belong to different words, in $V \times V$
(2)         IF ( REL$(S,S') \neq$ NULL and $S \neq S'$ THEN ADD edge $E(S,S')$ to $E$ and $E(S',S)$ to $E$
        FOR each word $W$ AND each of its word senses $S$ in $V$
(3)         INITIALIZE $P_S = 1/n(W)$ as the initial probability
```

```
(3a)    ASSIGN weight (1-$d$) to matrix element $M_{S,S}$
(3b)    COMPUTE $e(S)$ as the number of edges leaving $S$
        FOR each other word $W' \neq W$ in $WL$ AND each sense $S'$ of $W'$
(3c)        IF there is an edge between $S$ and $S'$ THEN ASSIGN Weight $d/e(S)$ to $M_{S,S'}$
            OTHERWISE ASSIGN 0 to $M_{S,S'}$
```

Score vertices in *G*

```
      REPEAT
          FOR each word $W$ AND each of its word senses $S$
(4)           INTIALIZE $Q_S$ to $P_S \times M_{S,S}$
              FOR each other word $W' \neq W$ in $WL$ AND each sense $S'$ of $W'$
(4a)              INCREMENT $Q_S$ by $P_{S'} \times M_{S',S}$
          FOR each word $W$, SUM $Q_S$ over $n(W)$ senses as $N_w$
          FOR each word $W$ AND each of its word senses $S$
(4b)          REPLACE $P_S$ by $Q_S/N_w$
      UNTIL probability $P_S$'s converge
```

Assign word sense

```
(5)  INITIALIZE $List$ as NULL
     FOR each word $W$ in $WL$
(6)      APPEND ($W$,$S$*) to $List$ where $P_{S*}$ is the maximum among senses of $W$
(7)  OUTPUT $List$
```

Figure 4. Algorithm for Graph-based Word Sense Assignment.

The algorithm for the best sense assignment S^* for W consists of three main parts: (1) construction of a weighted word sense graph; (2) sense scoring using the iterative Random Walk algorithm; and (3) word sense assignment.

In Step 1 of the algorithm, by referring to *SI*, we populate candidate $n(W)$ senses for each word W in the word list, *WL*, under the same semantic category as vertices in graph *G*. In *G*, directed edges $E(S,S')$ and $E(S',S)$ are built between vertex *S* and vertex *S'* if and only if there exists a semantic relation between the word sense *S* and *S'* in *SI*. Figure 5 shows an example of such a graph.

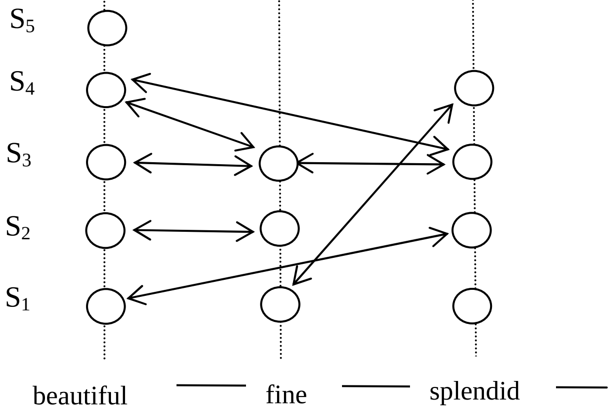


Figure 5. Sample graph built on the admissible word senses (vertical axis) for three words (horizontal axis) under the thesaurus category of "Goodness". Note that self-loop edges are omitted for simplicity.

We initialize the probability concerning the sense S of a word W, P_s, to $1/n(W)$, uniform distribution among the senses of W (Step (3)). For example, in Figure 5, the probability of the fourth sense of the word "beautiful" is initialized to 0.2. Then, we construct a matrix, whose element $M_{x,y}$ stands for the proportion of the probability P_x , that will be propagated to node y. Since $M_{x,y}$ may not be equal to $M_{,y,x}$, the edges in G are directed. In matrix M, we assign 1-d to $M_{x,x}$ where $x \in V$(Step (3a)) while the rest of the proportion (*i.e.*, d) is uniformly distributed among the outgoing edges of the node x (Step (3c)). Take the fourth sense (Node 4 for short) of the word "beautiful" and the third sense (Node 8 for short) of the word "fine" in Figure 5 for example. $\mathrm{M}_{4,8}$ is $d/2$ since there are two outgoing edges for Node 4. On the other hand, $\mathrm{M}_{8,4}$ is $d/3$ in that there are three edges leaving Node 8. d is the damping factor and was first introduced by PageRank (Brin & Page, 1998), a link analysis algorithm. The damping factor is usually set around 0.85, indicating that eighty-five percent of the probability of a node will be distributed to its outbound nodes.

In the second part of the algorithm, probabilities will be iteratively re-distributed among the senses of words until convergence of probabilities. For each sense S of a word W, first, (Step (4)) Q_s is assigned to $P_\mathrm{s} \times M_{s,s}$ (*i.e.*, some proportion, $M_{s,s}$, of the probability of P_s is propagated to the node s), then (Step (4a)) Q_s is incremented by $P_{\mathrm{s}'} \times M_{s',s}$, the ingoing probability propagation from node s', whenever there is an edge between s' and s. In Step (4b), we re-calculate the probability of the sense S, P_s, by dividing Q_s by $\sum_{s' \in sense(W)} Q_{S'}$, where S and S' are different word senses of the same word W and $sense(W)$ is the set of admissible senses of W in SI for the next iteration. $\sum_{s' \in sense(W)} Q_{S'}$, or N_w in the algorithm, is the normalization factor. The propagation of probabilities at each iteration in this graph-based algorithm, or Random Walk Algorithm, ensures that if a node is *semantically*[1] linked to another node with high probability, it will obtain quite a few probabilities from that node, indicating that this node may be important[2] in that probabilities converse and tend to aggregate in senses (*i.e.*, nodes) of words that are semantically related (*i.e.*, connected).

Finally, for each word, we identify the most probable sense and attach the sense to it (Step (6)). For instance, for the graph in Figure 6, the vertex on the vertical axis represented as the *sense #3* of "fine" will be selected as the best sense for "fine" under the thesaurus category "*Goodness*" with other entry words, such as, "lovely," "superb," "beautiful," and "splendid". The output of this stage is a set of linked word sense pairs (W, S*) that can be utilized to extend the coverage of thesauri via semantic relations in SI.

[1] Edges only exist when there is a semantic relation between vertices, or senses.

[2] As probable.

Theoretically, the method of PageRank (Brin & Page, 1998) distributes more probabilities or more scores through edges to well-connected nodes (*i.e.*, well-known web pages) in a network (*i.e.*, the Web). That is, more connected nodes tend to collect scores, in turn propagating comparatively more significant scores to their connected neighboring nodes. Consequently, the flow or re-distribution of probabilities or scores mostly would be confined to nodes in groups and the convergence of the probabilities over the network is to be expected normally. In this stage of our method, an edge is added if and only if there are some semantic relations, in the sense inventory, existing between two word senses (*e.g.*, one is the immediate hyponym/hypernym of the other), to differentiate semantically-related senses from those that are not. The PageRank-like algorithm in Figure 4 is exploited to determine the most well-connected or more semantically related (sense) group. Additionally, the senses in the group are assumed to be the most suitable senses of words for the given semantic category or semantic topic. This assumption is more likely to be correct if the number of given words in a category is big enough (it is usually easier to uniquely determine the sense of words given more words). Moreover, empirically, the number of iterations needed for probabilities to converge is less than ten (Usually, six is enough. It took only three iterations for words in Figure 6 to converge.); a quick scan of the results of this sense-assigning step reveals that the aforementioned assumption leads to satisfying sense analyses.

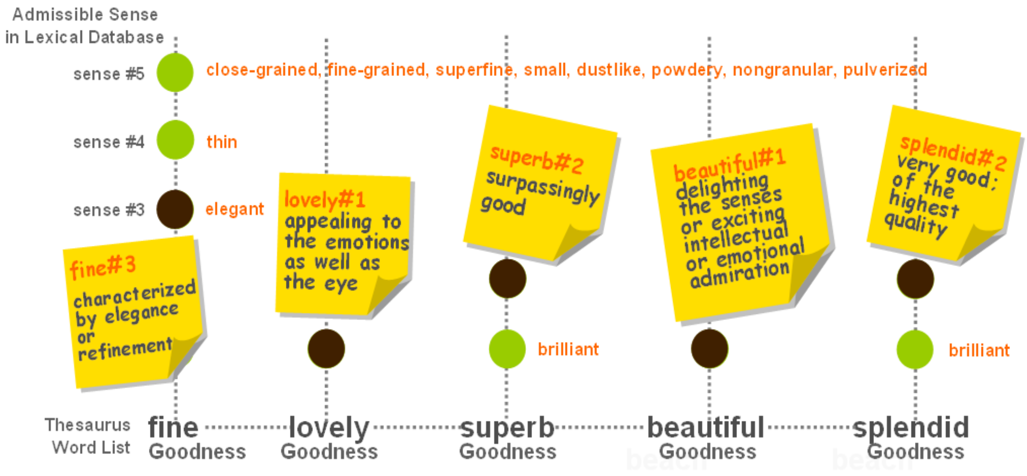

Admissible Sense
in Lexical Database
sense #5
close-grained, fine-grained, superfine, small, dustlike, powdery, nongranular, pulverized
sense #4
thin
sense #3
elegant
fine#3
characterized by elegance or refinement
lovely#1
appealing to the emotions as well as the eye
superb#2
surpassingly good
beautiful#1
delighting the senses or exciting intellectual or emotional admiration
splendid#2
very good; of the highest quality
brilliant
brilliant
Thesaurus
Word List
fine
Goodness
lovely
Goodness
superb
Goodness
beautiful
Goodness
splendid
Goodness

Figure 6. Highest scoring word sense in the stationary distributions for thesaurus word list under category "Goodness" assigned automatically by Random Walk on graph.

3.2.2 Extending the Coverage of Thesaurus

Automating the task of constructing a large-scale semantic knowledge base for semantic classification imposes a huge effort on the side of knowledge integration. Starting from a widespread computational lexical database, such as *WordNet*, overcomes the difficulties of building a knowledge base from scratch. In the second stage of the learning process (Step (2) in Figure 3), we attempt to broaden the limited thesaurus coverage in view of reducing encounters of unknown words in collocation label assignment in Section 3.3. The sense-annotated word lists generated as a result of the previous step are useful for enlarging and enriching the vocabulary of the thesaurus.

Take the sense-annotated result in Figure 6 for example. "Fine" with other adjective entries "beautiful," "lovely," "splendid," and "superb" under the semantic label "*Goodness*" is identified as belonging to the word sense *fine#3* "*characterized by elegance or refinement or accomplishment*" rather than other admissible senses (as shown in Table 1). After knowing the sense of the word "fine" under the semantic category "*Goodness*," we may now add its similar words via feasible semantic operators (as shown in Table 2) provided in the word sense inventory (*e.g*., *WordNet*). Its similar word, as suggested in Table 1 and 2, elegant#1 can be acquired by applying the operator "syn operator" on fine#3. Then, elegant#1 is incorporated into the knowledge base (*e.g*., *ISK*) under the semantic category of fine#3, "*Goodness*".

Table 1. Admissible senses for adjective "fine"

Sense Number	Definition	Example	Synsets of Synonym
fine #1	(being satisfactory or in satisfactory condition)	*"an all-right movie"; "everything's fine"; "the passengers were shaken up but are all right"; "dinner and the movies had been fine"; "things are okay"*	all right#1, o.k.#1, ok#1, okay#1, hunky-dory#1
fine #3	(characterized by elegance or refinement or accomplishment)	*"fine wine" ; "a fine gentleman"; "looking fine in her Easter suit"; "fine china and crystal"; "a fine violinist"*	elegant#1
fine #4	(thin in thickness or diameter)	*"a fine film of oil"; "fine hairs"; "read the fine print"*	thin#1

Table 2. Semantic relation operators for extending the coverage of thesaurus.

semantic relation operators	Description	Relations Hold for
syn operator	synonym sets for every word that are interchangeable in some context without changing the truth value of the preposition in which they are embedded	all words
sim operator	adjective synsets contained in adjective clusters	adjectives

In the end, by using semantic operators in lexical database (*e.g.*, *WordNet*), the coverage of the integrated semantic knowledge obtained from Step (1) in Figure 3 can be enlarged for assigning the semantic label of a collocation at run-time (Section 3.3).

3.2 Giving Thesaurus Structure to Collocation by Iterative Graphical Algorithms

Provided with the extended semantic knowledge obtained by following the learning process in Section 3.2, we build a thesaurus structure for the query results from online collocation reference tools. Figure 7 illustrates a thesaurus structure imposed on some adjective collocations (*i.e.*, "superb," "fine," "lovely," "beautiful," "splendid," *etc*.) of the word "beach" by our system.

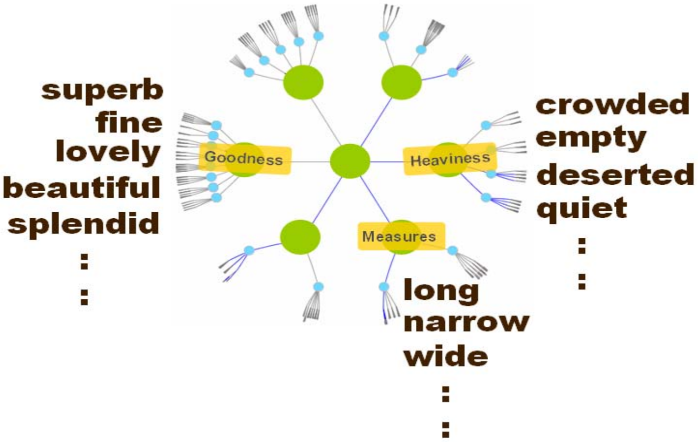


Figure 7. Sample adjective collocations of the word "beach" after being classified into some general-purpose semantic topics.

At run-time, we apply the Random Walk algorithm, which is very similar to the one in Figure 4, to automatically assign semantic labels to all collocations of a pivot word (*e.g.*, "beach") by exploiting semantic relatedness identified among these collocations. Once we know the semantic labels, or thesaurus categories, of the collocates, we partition them in groups according to their labels, which is helpful for dictionary look-up and for L2 learners to quickly find their desired collocations under some semantic meaning. The following depicts the semantic labeling procedure.

The input to this procedure is (1) a set of collocations, *Col*, for the query word *X*; (2) the integrated semantic knowledge (*i.e.*, *ISK*) from Section 3.2, {(*W*, *L*)} where a word *W* is semantically labeled as *L*. The output of this procedure is sets of collocations, each of which is classified under a semantic label and contains semantically-related collocations of the query word (see Figure 7).

At first, we construct a graph G={V,E} where a vertex in V represents a possible semantic category for a collocation C in *Col* and an edge in E represents a semantic relatedness holding between vertices. Note that we can look up possible semantic labels of a word from *ISK* and that edges in G are directed.

We use P_L to depict the probability of the candidate label, L, of a collocation in *Col*. Prior to the random-walking process, P_L is uniformly initialized over possible labels of a collocation. Once the matrix M, representing the proportions of probabilities to be propagated, is built, P_L will be iteratively changed, based upon current statistics, until convergence of probabilities. Recall that an element $M_{x,y}$ in the matrix will be set to 1-d if node x is equal to node y; will be set to $d/e(x)$ if x is different from y, there is an edge between x and y, and there are $e(x)$ edges leaving x; and will be set to zero otherwise. At each iteration, the probabilities of the candidate labels of a collocate sum to one, suggesting normalization is needed for each iteration as in the algorithm of word sense assignment in Figure 4.

Finally, we identify the most probable semantic label L^* for each collocate C, resulting in a list of (C, L^*). The procedure is designed to arrange given collocations in thesaurus categories with semantically related collocations therein, providing L2 learners with a thesaurus index for easy lookup or easy concept-grasping (see Figure 7 for an example).

4. Experimental Setting

4.1 Experimental Data

In our experiment, we applied the Random Walk Algorithm (in Section 3.2 and Section 3.3) to partition collocations into existing thesaurus categories, thus imposing a semantic structure on the raw data (*i.e.*, given collocations). In analysis of learners' collocation error patterns, verb-noun (V-N) and adjective-noun (A-N) collocations were found to be the most frequent error patterns (Liu, 2002; Chen, 2002). Hence, for our experiments and evaluation, we focused our attention particularly on V-N and A-N collocations.

Recall that our classification model starts with a thesaurus consisting of lists of semantically related words and extends the thesaurus using sense labeling in Section 3.2.1 and semantic operators in the word sense inventory in Section 3.2.2. The extended semantic knowledge provides collocates with topic labels for semantic classification of interest. Two kinds of resources required in our experiment to obtain the extended knowledge base are described below.

4.1.1 Data Source 1: A Thesaurus

We used *Longman Lexicon of Contemporary English* (*LLOCE* for short) as our thesaurus of semantic categories (*i.e.*, *TC*). *LLOCE* contains 15,000 distinct entries for all open-class words, providing semantic fields of a pragmatic, everyday common sense index for easy reference. The words in *LLOCE* are organized into approximately 2,500 semantic word sets. These sets are divided into 129 semantic categories and further organized as 14 semantic fields. Thus, the semantic field, category, and word set in *LLOCE* constitute a three-level hierarchy, in which each semantic field contains 7 to 12 categories and each category contains 10 to 50 sets of semantic related words. The *LLOCE* is based on coarse, topical semantic classes, making them more appropriate for WSD than other finer-grained lexica. Alternatively, *Roget's Thesaurus* can be used as the thesaurus.

4.1.2 Data Source 2: A Word Sense Inventory

For our experiments, we need comprehensive coverage of word senses. Word senses can be obtained easily from any definitive record of the English language (*e.g*. an English dictionary, encyclopedia or thesaurus). We used *WordNet 3.0* as our sense inventory. It is a broad-coverage, machine-readable lexical database, publicly available in parsed form (Fellbaum, 1998) and consists of 212,557 sense entries for open-class words, including nouns, verbs, adjectives, and adverbs. *WordNet* is organized by the synonymous sets, or synsets, and provides semantic operators to act upon its synsets.

4.2 Experimental Configurations

Given the aforementioned two data sources, we first integrate them into one then broaden the vocabulary of the thesaurus, the basis knowledge for assigning semantic labels to collocations.

4.2.1 Step 1: Integrating Semantic Knowledge

For each semantic topic in *LLOCE*, we attach word senses to its constituent words based on semantic coherence (within a topic) and semantic relations created by lexicographers from *WordNet*. The integrated semantic knowledge can help interpret a word by providing information on its word sense and its corresponding semantic label.

Recall that, to incorporate senses into words with semantic topics, our model applies the Random Walk Algorithm on a weighted directed graph whose vertices (word senses) and edges (semantic relations) are extracted from and are based on *LLOCE* and *WordNet 3.0*. All edges are drawn and weighted to represent the magnitudes of semantic relatedness among word senses. See Table 3 for the relations (or semantic operators) existing in edges in our experiment.

Table 3. Available semantic relations.

Relations	Semantic Relations for Word Meanings	Relations Hold for
syn	synonym sets for every word that are interchangeable in some context without changing the truth value of the preposition in which they are embedded	all words
hyp	hypernym/hyponym (superordinate/subordinate) relations between synonym sets	nouns verbs
vgp	verb synsets that are similar in meaning and should be grouped together when displayed in response to a grouped synset search.	verbs
sim	adjective synsets contained in adjective clusters	adjectives
der	words that have the same root form and are semantically related	all words

4.2.2 Step 2: Extending Semantic Knowledge

Based on the senses mapped to words with semantic labels (via the graph-based sense assignment algorithm), we further utilize the semantic operators in *WordNet* (*i.e.*, *SI*) to add new words into *LLOCE* (*i.e.*, *TC*). Depending on the part-of-speech (*i.e.*, noun, adjective, or verb) of the word at hand, various kinds of semantic relation operators (see Table 3) are available for enriching the vocabulary of the integrated semantic knowledge (*i.e.*, *ISK*) of *WordNet* and *LLOCE*. In the experiment, using the *syn* operator alone broadened the vocabulary size of *ISK* to a size more than twice as large as that of the thesaurus *LLOCE* (*i.e.*, 39,000 vs. 15,000).

4.3 Test Data

We used a collection of 859 V-N and A-N collocation pairs for testing. These collocations were obtained from the website: *JustTheWord* (http://193.133.140.102/JustTheWord/). *JustTheWord* clusters collocates into sets without any explicit semantic label. We will compare its clustering performance with our model's performance in Section 5.

In the experiment, we evaluated semantic classification of three[3] types of collocation pairs: (1) A-N pairs and clustering over the **<u>adjectives</u>** (**<u>A</u>**-N), (2) V-N pairs and clustering over the **<u>verbs</u>** (**<u>V</u>**-N), and (3) V-N pairs and clustering over the **<u>nouns</u>** (V-**<u>N</u>**). For each type, we selected five pivot words with varying levels of abstractness for L2 learners and extracted a subset of their respective collocations from *JustTheWord*, leading to a test data set of 859 collocation pairs. Table 4 shows the number of the collocations for each pivot of each collocation type. In total, 307 collocates were extracted for **<u>A</u>**-N, 184 for **<u>V</u>**-N, and 368 for

[3] We do not consider the case of A-<u>N</u> in that, usually, various nouns can follow an adjective.

V-**N**.

To appropriately select our testing pairs from *JustTheWord*, we were guided by research into L2 learners' and dictionary users' needs and skills for second language learning, especially taking account the meanings of complex words with many collocates (Tono, 1992; Rundell, 2002). The pivot words we selected for testing are words that have many respective collocations and are shown in worth-noting boxes in *Macmillan English Dictionary for Advance Learners* [ISBN 0-333-95786-5] (First edition, henceforth *MEDAL*).

Table 4. Statistics of our testing collocation pairs.

collocation type	pivot word	some collocations	count
A-N (N=pivot)	advice	helpful, dietary, impartial, free	36
	attitude	healthy, moral, aggressive, right	49
	description	clinical, excellent, fair, precise	47
	effect	serious, inevitable, possible, sound	114
	impact	dramatic, negative, powerful, severe	61
V-N (N=pivot)	balance	strike, maintain, achieve, tilt, tip	29
	disease	cure, combat, carry, transmit, carry	21
	issue	settle, clarify, identify, remain, avoid	38
	plan	propose, submit, accept, involve	54
	relationship	forge, alter, develop, damage, form	42
V-**N** (V=pivot)	deserve	blame, support, title, thanks, honor	51
	express	love, anger, fear, personality, doubt	82
	fight	disease, war, , enemy, cancer, duel	24
	hold	funeral, presidency, hope, knife	151
	influence	health, government, opinion, price	60

5. Results and Discussions

Two pertinent sides were addressed for the evaluation of our results. The first was whether such a model for a thesaurus-based semantic classification could generate collocation clusters correlating with human word meaning similarities to a significant extent. Second, supposing it could, would its results of semantic label assignment lead to easy dictionary lookup or better collocation understanding and production? In the following sections, two evaluation metrics are described to respectively examine our results in these two aspects, that is, the accuracy of our collocation clusters and the helpfulness of our labels in terms of language learning.

5.1 Performance Evaluation for Semantic Clusters

Traditional cluster evaluation (Salton, 1989) might not be suited to assess our model, where we aim to facilitate collocation referencing and help learners improve their collocation production. Hence, to evaluate the performance of our clustering results, an evaluation sheet made up of test items, resembling synonym test items of the Test of English as a Foreign Language (TOEFL), was automatically generated for human judgment. Landauer and Dumais (1997) first proposed using the synonym test items of TOEFL as an evaluation method for semantic similarity. Fewer fully automatic methods of a knowledge acquisition evaluation, *i.e.*, ones that do not depend on knowledge being entered by a human, have been capable of performing well on a full scale test used to measure semantic similarity. A test item provided by Landauer (1997, as cited in Padó & Lapata, 2007) is shown below where "crossroads" is the synonym for "intersection" in the context.

You will find the office at the main **intersection**.

(a) place (b) crossroads (c) roundabout (d) building

As to our experiment, we evaluated the semantic relatedness among collocation clusters according to the above-mentioned TOEFL benchmark by setting up test items out of our clustering results. Then, human judges performed a decision task similar to TOEFL test takers: deciding which one of the four alternatives was synonymous with the target word. A sample question is shown below where "rocky" is clearly the most similar word for "sandy" given the pivot word "beach".

***sand*y** beach

(a) long (b) rocky (c)super (d)narrow

There were 150 multiple-choice questions randomly constructed to test the accuracy of our clusters, 50 questions for each collocation types (*i.e.*, **A**-N, **V**-N, and V-**N**) and 10 for each of collocation pairs. In order to evaluate the degree to which our model achieved production of good clusters, two judges were asked to choose the most appropriate answer. More than one answer was allowed if the judges found some of the distractors in the test items to be plausible answers. Moreover, the judges were allowed not to choose any of the alternatives given if they thought no satisfactory answer was provided. Table 5 shows the performance of collocation clusters generated by *JustTheWord* and the proposed system. As suggested in the table, our model achieved significantly higher precision and recall in comparison with our baseline, *JustTheWord*.

Table 5. Precision and recall of two systems

Results / System	Judge 1		Judge 2		Inter-Judge Agreement
	Precision	**Recall**	**Precision**	**Recall**	
Ours	.79	.71	.73	.67	.82
JustTheWord	.57	.58	.57	.59	

With high inter-judge agreement (*i.e.*, 0.82), the influence of human judges' subjectivity on the performance evaluation of collocation clusters is not that severe and it is modest to say that our model's clustering results are thought to be better than the baseline's across human judges.

5.2 Conformity of Semantic Labels

The second evaluation task focused on whether the semantic labels would facilitate users scanning the collocation entries quickly and finding the desired concept of the collocations. The evaluation is aimed at examining the extent to which semantic labels are useful, and to what degree of reliability.

Two native speakers were asked to grade half of the labeled collocations randomly selected from our classifying results (all test data considered). A three-point rubric is used to evaluate the effectiveness, or usefulness, of the given semantic labels in terms of navigating users to the desired collocates. The three types of rubric points with their descriptions are: three points for those collocations with effective semantic labels in navigation in a collocation reference tool, two points for those with somewhat helpful assigned labels, and one point for those with misleading labels.

Table 5 shows that 77% of the semantic labels assigned as a reference guide have been judged as adequate in terms of guiding a user finding a desired collocation in a collocation learning tool and that our classification model provably yields productive performance of semantic labeling of collocates to be used to assist language learners. The results justify the thought that the move towards semantic classification of collocations is of probative value.

Table 6 shows that 76% of the semantic labels assigned as a reference guide were judged adequate in terms of guiding users to find a desired collocation in a collocation learning tool, and this suggests that our classification model yielded promising performance in semantically labeling collocates further to be used to assist language learners. The results justify that the move towards semantic classification of collocations is of probative value.

Table 6. Performance evaluation for assigning semantic labels as a reference guide

	Judge 1	**Judge 2**
Ours	.77	.75
JustTheWord	Not available	Not available

6. Conclusion and Future Work

The research seeks to create a thesaurus-based semantic classifier within a collocation reference tool without meaning access indices. We describe a thesaurus-based semantic classification for a semantic grouping of collocates with a pivot word. The construction of a collocation thesaurus is meant to enhance L2 learners' collocation production. Our classification model is based on two graph-based Random Walk Algorithms (*i.e.*, word sense assignment and semantic label assignment) to categorize collocations into semantically-related groups for easy dictionary lookup and collocation understanding and production. The limited vocabulary size of the semantic thesaurus is dealt with using the sense information and the semantic operators in the word sense inventory, *WordNet*. The evaluation shows that the thesaurus structure imposed by our model for an existing computational collocation reference tool is quite accurate and is helpful for users to navigate the collocations of a pivot word.

Many avenues exist for future research and improvement of our system. For example, semantic relations existing between word senses may take on different weights in that some may be more informative than others in determining semantic similarities. Another interesting direction to explore is to see if our model can benefit from other thesauri with semantic labels.

References

Benson, M. (1985). Collocations and Idioms. In *R. Ilson (Ed.), Dictionaries, Lexicography and Language Learning* (ELT Documents 120; Oxford: Pergamon), 61-68.

Béjoint, H. (1994). Tradition and Innovation in Modern English Dictionaries. Oxford: *Clarendon Press*.

Brants, T. & Franz, A. (2006). Web 1T 5-gram corpus version 1.1. Technical report, Google Research.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW Conference*.

Chen, Y. (2004). A corpus-based analysis of collocational errors in EFL Taiwanese High School students' compositions. California State University, San Bernardino. June.

Pantel, P. & Chklovski, T. (2004). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 33-40.

Chen, J.-N. & Chang, J. S. (1998). Topical clustering of MRD senses based on information retrieval techniques, *Computational Linguistics*, 24(1), March 1998.

Downing, S. M., Baranowski, R. A. , Grosso, L.J., & Norcini, J. J. (1995). Item type and cognitive ability measured: the validity evidence for multiple true-false items in medical specialty certification. *Appl Meas Educ*, 8, 189-199.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* (*JASIS*), 41(6), 391-407.

Firth, J. R. (1957). The Semantics of Linguistics Science. Papers in linguistics 1934-1951. London: *Oxford University Press*.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Hall, G. (1994). Review of The Lexical Approach: The State of ELT and a Way Forward, by Michael Lewis. *ELT Journal,* 44, 48.

Heimlich, J. E. & Pittelman, S. D. (1986). Semantic mapping: Classroom applications. Newark, DE: International Reading Association.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Meeting of the Association for Computational Linguistics*, 268-275.

Hatzivassiloglou, V., & McKeown, K. R. (1993). Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st ACL*, 172-182.

Jian, J.-Y., Chang, Y.-C., & Chang, J. S. (2004). TANGO: Bilingual Collocational Concordancer, *Post & demo in ACL* 2004, Barcelona.

Johnson, D. D., & Pearson, P. D. (1984). *Teaching reading vocabulary*. New York: Holt, Rinehart & Winston.

Kilgarriff, A. (1997). I Don’t Believe in Word Senses, In: *Computers and the Humanities*. 31(2), 91-113(23).

Kilgarriff, A. & Tugwell, D. (2001). “WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography”. In *Proceedings of COLLOCATION: Computational Extraction, Analysis and Exploitation workshop*, 32-38.

Kemp, J. E., Morrison, G. R., & Ross, S. M. (1994). Developing evaluation instruments. In: *Designing Effective Instruction*. New York, NY: MacMillan College Publishing Company, 180-213.

Lewis, M. (1997). Implementing the lexical approach. Hove, England: *Language Teaching Publications*.

Lewis, M. (2000). Language in the Lexical Approach. In. M. Lewis (ed.) Teaching Collocation: Further development in the Lexical Approach. London, Language Teaching Publications.

Landauer, T. & Dumais, S. T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.

Lesk, M. E. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC ’86*, 24-26.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Meeting of the Association for Computational Linguistics*, 64-71.

Liu, L. E. (2002). A corpus-based lexical semantic investigation of vernb-noun miscollocations in Taiwan learners' English. *Tamkang University*, Taipei, January.

Morris, J. & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 21-48.

Nattinger, J. R., & DeCarrico, J. S. (1992). Lexical Phrases and Language Learning. Oxford: *Oxford University Press*.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223-242.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge: Cambridge Press.

Nirenburg, S. & Raskin, V. (1987). The subworld concept lexicon and the lexicon management system, *Computational Linguistics*, v. 13, December 1987.

Nastase, V. & Szpakowicz, S. (2003). Exploring noun–modifier semantic relations. In *Fifth International Workshop on Computational Semantics* (*IWCS-5*), 285-301.

Padó, S. & Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2), 161-199.

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1155-159.

Readence, J. E., & Searfoss, L.W. (1986). Teaching strategies for vocabulary development. In E. K. Dishner, T. W. Bean, J. E. Readence, & D. W. Moore (Eds.), *Reading in the content areas: Improving classroom instruction* (2nd ed., pp. 183-188). Dubuque, IA: Kendall/ Hunt.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337-354.

Scholfield, P. (1982). Using the English dictionary for comprehension. *TESOL Quarterly*, 16, 185-194.

Salton, G. (1989). Automatic Text Processing: The transformation, analysis, and retrieval of information by computer. *Addidon-Wesley*.

Sinatra, R., Beaudry, I., Pizzo, I., & Geishart, G. (1994). Using a computer-based semantic mapping, reading and writing approach with at-risk fourth graders. *Journal of Computing in Childhood Education*, 5, 93-112.

Tono, Y. (1984). On the Dictionary User's Reference Skills. Unpublished B.Ed. Thesis. Tokyo: *Tokyo Gakugei University*.

Tono, Y. (1992). The Effect of Menus on EFL Learners' Look-up Processes. LEXICOS 2 (AFRILEX Series) Stellenbosch: *Buro Van de Watt*.

Taba, H. (1967). Teacher's handbook for elementary social studies. Reading, MA: Addison-Wesley.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417-424.

Turney, P. D. (2006). Similarity of Semantic Relations. *Computational Linguistics*, 32(3), 379-416.

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 3, September 2009, pp. 281-304

Automatic Recognition of Cantonese-English Code-Mixing Speech

Joyce Y. C. Chan[*], Houwei Cao[*], P. C. Ching[*], and Tan Lee[*]

Abstract

Code-mixing is a common phenomenon in bilingual societies. It refers to the intra-sentential switching of two different languages in a spoken utterance. This paper presents the first study on automatic recognition of Cantonese-English code-mixing speech, which is common in Hong Kong. This study starts with the design and compilation of code-mixing speech and text corpora. The problems of acoustic modeling, language modeling, and language boundary detection are investigated. Subsequently, a large-vocabulary code-mixing speech recognition system is developed based on a two-pass decoding algorithm. For acoustic modeling, it is shown that cross-lingual acoustic models are more appropriate than language-dependent models. The language models being used are character tri-grams, in which the embedded English words are grouped into a small number of classes. Language boundary detection is done either by exploiting the phonological and lexical differences between the two languages or is done based on the result of cross-lingual speech recognition. The language boundary information is used to re-score the hypothesized syllables or words in the decoding process. The proposed code-mixing speech recognition system attains the accuracies of 56.4% and 53.0% for the Cantonese syllables and English words in code-mixing utterances.

Keywords: Automatic Speech Recognition, Code-mixing, Acoustic Modeling, Language Modeling

1. Introduction

Code-switching and code-mixing are common phenomena in bilingual societies. According to John Gumperz (Gumperz, 1982), the definition of code-switching is “the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical

[*] Department of Electronic Engineering, The Chinese University of Hong Kong
E-mail: {ycchan, hwcao, pcching, tanlee}@ee.cuhk.edu.hk

systems or sub-systems". Different combinations of languages are found in code-switching, for examples, Spanish-English in United States, German-Italian and French-Italian in Switzerland, and Hebrew-English in Israel (Auer, 1998). In Taiwan, code-switching between Chinese dialects, namely Mandarin and Taiwanese, has become common in recent years (Chen, 2004). Hong Kong is an international city where many people, especially the younger generation, are Cantonese and English bilinguals. English words are frequently embedded into spoken Cantonese. The switching of language tends to be intra-sentential, and it rarely involves linguistic units above the clause level. Hence, the term code-mixing is usually preferred (Li, 2000). In this case, Cantonese is the primary language, also known as the matrix language, and English is the secondary language, usually referred to as the embedded language (Halmari, 1997).

Automatic speech recognition (ASR) is one of the key technologies in spoken language processing. An ASR system converts an input speech waveform into a sequence of words. Recently, ASR for multilingual applications has attracted great interest (Schultz & Kirchhoff, 2006). In state-of-the-art ASR systems, the input speech is assumed to contain only one language and the language identity is given. These systems are not able to handle code-mixing speech, which differs significantly from monolingual speech spoken by native speakers. This calls for special consideration in the design of acoustic models, lexical and language models, and in the decoding algorithm.

There have been two different approaches to code-switching or code-mixing speech recognition (Lyu *et al.,* 2006; Chan *et al.,* 2006). The first approach involves a language boundary detection (LBD) algorithm that divides the input utterance into language-homogeneous segments. The language identity of each segment is determined, and the respective monolingual speech recognizer is applied. LBD for mixed-language utterances was studied by Wu *et al.* (2006) and Chan *et al.* (2004). Language-specific phonological and acoustic properties were used as the primary cues to identify the languages. The second approach aims to develop a cross-lingual speech recognition system, which can handle multiple languages in a single utterance. The acoustic models, language models, and pronunciation dictionary are designed to be multi-lingual and cover all languages concerned. In Lyu *et al.* (2006), automatic recognition of Mandarin-Taiwanese code-switching speech was investigated. It was found that Mandarin and Taiwanese, both of which are Chinese dialects, share a large percentage of lexicon items. Their grammar was also assumed to be similar. A one-pass recognition algorithm was developed using a character-based search net. It was shown that the one-pass approach outperforms LBD-based multi-pass approaches. In You *et al.* (2004), a mixed-lingual keyword spotting system was developed for auto-attendant applications. The keywords to be detected could be in either English or Chinese.

This paper presents a study on automatic speech recognition of Cantonese-English code-mixing speech. Part of the work was reported in Chan *et al.* (2006). Our study covers all components of an ASR system, including acoustic models, language models, pronunciation dictionary, and search algorithm. Different approaches to LBD are also investigated. By understanding the linguistic properties of monolingual Cantonese and English, as well as code-mixing speech, the major difficulties in code-mixing speech recognition are revealed and possible solutions are suggested. We propose a two-pass recognition system, in which the acoustic and linguistic knowledge sources are integrated with language boundary information. Simulation experiments are carried out to evaluate the performance of the whole system as well as individual system components.

2. Difficulties in Code-mixing Speech Recognition

2.1 Linguistic Properties of Cantonese and English

Cantonese is a Chinese dialect. It is spoken by tens of millions of people in the provinces of Guangdong, Guangxi, Hong Kong, and Macau. A Chinese word in its written form is composed of a sequence of characters. In Cantonese, each Chinese character is pronounced as a monosyllable carrying a specific lexical tone (Ching *et al.,* 2006). English is one of the most popular languages in the world. An English word is written as a sequence of letters. In spoken form, each word may consist of several syllables, some of which are designated to be stressed. Table 1 shows a pair of example words in Cantonese and English.

Table 1. Examples of Cantonese and English words in written and spoken format.

Written (orthographic transcription)	Spoken (phonetic transcription)
產生	/ts^{h} a n/ /s ɐ ŋ/
produce	/p r ə ˈd j uː s/

Syllables can be divided into smaller units, namely consonants (C) and vowels (V). Cantonese syllables take the structures of V, CV, CVC, or VC (Ching *et al.,* 1994). If tonal difference is not considered, the number of distinct Cantonese syllables is around 600 (Ching *et al.,* 2006). The syllable structure in English is more complicated than that in Cantonese. Although many English syllables share the same canonical forms as given above, there also exist combinations like CCV, VCC, CCCV, and CCCVCC (Wester, 2003), which are not found in Cantonese.

There are 22 consonants and 22 vowels (including diphthongs) in Cantonese, and 24 consonants and 14 vowels in American English (Ching *et al.,* 1994; Ladefoged, 1999). Table 2 lists the IPA (International Phonetic Alphabet) symbols of these phonemes. Some of the phonemes in the two languages are labeled with the same IPA symbols by phoneticians, meaning that they are phonetically very close. Some of the other phonemes are also considered to be very similar although they are labeled differently in the two languages, *e.g.*, /au/ in Cantonese and /aʊ/ in English.

Table 2. Phonemes of Cantonese and English. The phonemes that are labeled with the same IPA symbols in both Cantonese and English are listed first and boldfaced.

Cantonese phonemes	
IPA symbol	Example
p	［p a］（爸）
m	［m a］（媽）
f	［f a］（花）
t	［t a］（打）
tʃ	［tʃ y］（朱）
n	［n a］（拿）
s	［s a］（沙）
ʃ	［ʃ y］（書）
l	［l a］（啦）
j	［j ɐu］（憂）
k	［k a］（加）
ŋ	［p^h a ŋ］（烹）
w	［w a］ 蛙
h	［h a］（蝦）
ɪ	［s ɪ k］（色）
i	［s i］（絲）
ɛ	［s ɛ］（借）
ʊ	［s ʊ ŋ］（鬆）
u	［f u］（夫）
p^h	［p^h a］（扒）
t^h	［t^h a］（他）
ts	［ts i］（之）
ts^h	［ts^h i］（痴）
$tʃ^h$	［$tʃ^h$ y］（處）
k^h	[k^h a] (卡)
k^w	［k^w a］（瓜）
k^{wh}	［k^{wh} a］（誇）
y	［ʃ y］（書）
œ	［h œ］（靴）
a	［s a］（沙）
ɐ	［s ɐ p］（濕）
ɵ	［s ɵ t］（恤）
ɔ	［s ɔ］（梳）
ei	［h ei］（稀）
ɛu	［t ɛu］（投）
ai	［w ai］（威）
ɵy	［s ɵy］（衰）
ɐi	［s ɐi］（西）
ui	［f ui］（灰）
iu	［s iu］（燒）
ɐu	［s ɐu］（收）
au	［s au］（筲）
ɔi	［s ɔi］（鰓）
ou	［s ou］（鬚）

English phonemes	
IPA symbol	Example
p	[p aɪ] (pie)
m	[m aɪ] (my)
f	[f l aɪ] (fly)
t	[t aɪ] (tie)
tʃ	[tʃ ɪ n] (Chin)
n	[n ɛ t] (net)
s	[s æ t] (sat)
ʃ	[ʃ aɪ] (shy)
l	[l aɪ] (lie)
j	[j u] (you)
k	[k aɪ t] (kite)
ŋ	[h æ ŋ] (hang)
w	[w aɪ] (why)
h	[h aɪ] (high)
ɪ	[b ɪ d] (bid)
i	[b i t] (beat)
ɛ	[b ɛ d] (bed)
ʊ	[g ʊ d] (good)
u	[b u t] (boot)
b	[b aɪ] (buy)
v	[v aɪ] (vie)
θ	[θ ɪ ŋ] (thing)
ð	[ð e ɪ] (they)
d	[d aɪ] (die)
z	[z u] (zoo)
ɹ	[ɹ ɛ n t] (rent)
dʒ	[p e ɪ dʒ] (page)
ʒ	[æ ʒ ɚ] (azure)
g	[g aɪ] (guy)
e	[b e ɪ t] (bait)
æ	[b æ d] (bad)
ɚ	[b ɚ d] (bird)
o	[b o t] (boat)
ɑ	[p ɑ d] (pod)
ʌ	[b ʌ d] (bud)
aʊ	[k aʊ] (cow)
aɪ	[b aɪ] (buy)
ɔɪ	[b ɔɪ] (boy)

In this section, we use IPA symbols to facilitate an intuitive comparison between Cantonese and English. Language-specific phonemic symbols have been commonly used in monolingual ASR research, for examples, Pinyin for Mandarin, Jyut-Ping for Cantonese (LSHK, 1997), and ARPABET for American English (Shoup, 1980). In Section 4, where phoneme-based acoustic modeling is discussed, we will use Jyut-Ping and ARPABET for monolingual Cantonese and English respectively, and a combination of them for code-mixing speech.

2.2 Properties of Cantonese-English Code-mixing Speech

Table 3 gives an example of Cantonese-English code-mixing sentence spoken in Hong Kong. It contains an English segment with one word. In this case, the English word is used as a substitute for its Chinese equivalent. The grammatical structure is totally that of Cantonese. In our application, the mother tongue of the speaker is Cantonese, *i.e.*, the matrix language. It is inevitable that the embedded English words carry Cantonese accent to certain extent. In many cases, the syllable structure of an English word changes to follow the structure of legitimate Cantonese syllables (Li, 1996). Such changes usually involve phone insertions or deletions. For example, the second consonant in a CCVC syllable of English may be softened, *e.g.*, the word “plan” in the example of Table 3 is pronounced as /**p æ n**/ instead of /**p l æ n**/ by many Cantonese speakers. A monosyllabic word with the CVCC structure may become disyllabic by inserting a vowel at the end, *e.g.*, /**f æ n z**/ (“fans”) becoming /**f æ n s ɪ**/. It is also noted that the final stop consonant in an English word tends to be softened or dropped, *e.g.*, /**t ɛ s t**/ (“test”) becoming /**t ɛ s**/. This is related to the fact that the stop coda of a Cantonese syllable is unreleased (Ching *et al.,* 2006). In addition to phone insertion and deletion, there also exist phone changes in Cantonese-accented English. That is, an English phoneme that is not found in Cantonese is replaced by a Cantonese phoneme that people consider to sound similar. For example, /**ɵ r i**/ (“three”) becomes /**f r i**/ in Cantonese-accented English. Cantonese speakers in Hong Kong sometimes create a Cantonese pronunciation for an English word. For example, the word “file” (/**f aɪ l**/) is transliterated as /**f aɪ l o**/ (快佬 in written form). It is not a straightforward decision whether such a word should be treated as English or Cantonese. This is known as “lexical borrowing” (Chan, 1992).

In conclusion, English words in a code-mixing utterance must not be treated as being the same as those in a monolingual utterance from a native English speaker. For the design of ASR systems, special considerations are needed in acoustic modeling and lexicon construction.

Code-mixing occurs less frequently in read-style speech than in casual conversational speech. There exist many pronunciation variations in casual Cantonese speech, especially when the speaking rate is fast. Speakers may not follow strictly the pronunciations as specified in a standard dictionary. In the example of Table 3, the initial consonant /**n**/ of the first syllable is commonly pronounced as /**l**/ by the younger generation. Syllable fusion is often seen in fast speech, *i.e.*, the initial consonant of the second syllable of a disyllabic word tends to be omitted or changed (Kam, 2003; Wong, 2004).

Table 3. An example of a Cantonese-English code-mixing sentence

Code-mixing speech				
你哋 You (plural)	plan plan	咗 already	行程 schedule	未 ? or not
Transcription according to standard pronunciation dictionary				
/nei/ /tei/	/p l æ n/	/ts ɔ/	/h ɐŋ/ /ts^h ɪ ŋ/	/m ei/
Transcription according to typical pronunciation in code-mixing speech				
/lei/ /tei/	/p æ n/	/ts ɔ/	/h ɐŋ/ /ts^h ɪ ŋ/	/m ei/
English translation				
Have you planned your schedule already?				

2.3 Problems and Difficulties in Code-mixing Speech Recognition

Large-vocabulary continuous speech recognition (LVCSR) systems deal with fluently spoken speech with a vocabulary of thousands of words or more (Gauvain & Lamel, 2000). As shown in Figure 1, the key components of a state-of-the-art LVCSR system are acoustic models, pronunciation dictionary, and language models (Huang *et al.*, 2001). The acoustic models are a set of hidden Markov models (HMMs) that characterize the statistical variation of the input speech features. Each HMM represents a specific sub-word unit such as a phoneme. The pronunciation dictionary and language models are used to define and constrain the ways in which the sub-word units can be concatenated to form words and sentences.

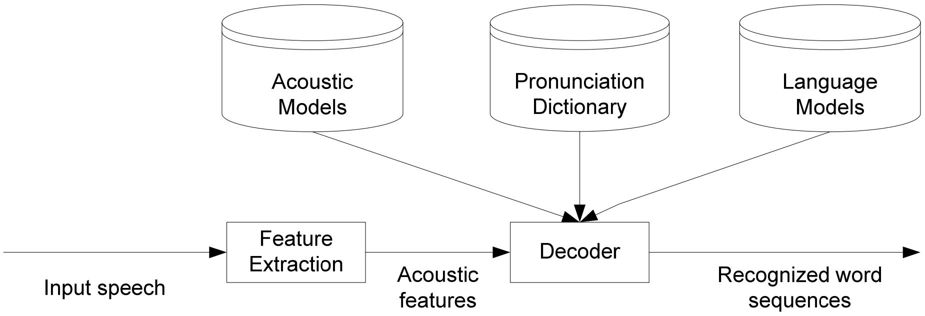


Figure 1. The flow diagram of an LVCSR system

For code-mixing speech recognition, the input utterance contains both Cantonese and English. Thus, the acoustic models are expected to cover all possible phonemes in the two languages. There are two possible approaches: (1) monolingual modeling with two separate sets of language-specific models; (2) cross-lingual modeling with some of the phoneme models shared between the two languages. Monolingual modeling has the advantage of preserving the language-specific characteristics and is most effective for monolingual speech from native speakers (Schultz & Waibel, 1998). In code-mixing speech where the English words are Cantonese-accented, an English phoneme tends to resemble or even become identical to a Cantonese counterpart. In this case, we may treat them as the same phoneme and establish a cross-lingual model to represent it. As shown in Table 2, Cantonese and English have a number of phonemes that are phonetically identical or similar to each other. The degree of similarity varies. In principle, cross-lingual modeling can be applied to those highly similar phonemes, while language-specific models would be more appropriate if the phonetic variation is relatively large. In Section 4, we are going to compare the effectiveness of cross-lingual and mono-lingual acoustic modeling and try to establish an optimal phoneme set for code-mixing speech recognition.

The pronunciation dictionary for code-mixing speech recognition is a mixture of English and Cantonese words. Each word may correspond to multiple pronunciations, which are represented in the form of phoneme sequences. Due to the effect of the Cantonese accent, the English words in code-mixing speech are subject to severe pronunciation variation as compared to those in standard English by native speakers. It is essential to reflect such variation in the pronunciation dictionary. On the other hand, as discussed in Section 2.2, the common pronunciation variations in spoken Cantonese should also be included.

In our application, the most common type of code-mixing is where one or more Cantonese words in the utterance being replaced by the English equivalent (Chan, 1992). The grammatical structure of code-mixing sentences is based largely on that of monolingual Cantonese. Word n-gram is by far the most commonly used technique for language modeling in LVCSR. To train a set of good n-gram models, a large number of spoken materials in computer-processable text format are needed. This presents a great challenge to our research since it is difficult in practice to find such materials for code-mixing speech. For the training of acoustic models, we need a large amount of code-mixing speech data. Development of speech and text corpora is therefore an important part of our work.

3. Development of Code-mixing Speech Corpus

In this section, the design, collection, and annotation of a Cantonese-English code-mixing speech corpus, named CUMIX, are described (Chan *et al.,* 2005). CUMIX is intended mainly for acoustic modeling for large-vocabulary speech recognition.

3.1 Corpus Design

There are three different types of utterances in CUMIX:

1. Cantonese-English code-mixing utterances (CM)
2. Monolingual Cantonese utterances (MC)
3. Monolingual English words and phrases (ME)

The CM utterances represent the typical code-mixing speech being dealt with in our application. There are practical difficulties in designing the content of code-mixing sentences. This is because spoken Cantonese is considered as a colloquial language that mainstream written publications do not use. Although the grammar of spoken Cantonese is similar to that of standard written Chinese, the lexical preference is quite different. An example pair of spoken Cantonese and written Cantonese sentences is shown in Table 4. Spoken Cantonese rarely appears in published text materials. Thus, text materials that involve code-mixing of spoken Cantonese and English are very limited.

Table 4. Comparison of spoken Cantonese and standard Chinese

Written Chinese:	你	吃過	午飯	了嗎?
Spoken Cantonese:	你	食咗	晏	未?
English translation (word by word):	You	eaten	lunch	or not?
English translation (whole sentence):	Have you had lunch?			

The design of CM sentences in CUMIX was based on a few local newspapers and online resources, including newsgroups and online diaries. We also consulted previous linguistic studies on Cantonese-English code-mixing. In Chan (1992), about 600 code-mixing sentences were analyzed. In 80% of the cases, the English segment contains a single word. The percentage distribution of nouns, verbs, and adjectives/adverbs are 43%, 24%, and 13%, respectively. There are very few cases involving prepositions and conjunctions. We try to follow these distributions in our corpus design.

A total of 3167 distinct code-mixing sentences were manually designed. Each sentence has exactly one English segment, which may contain one or more words. There are a total of 1097 distinct English segments. Each of them may appear more than once in the corpus, and if it does, the Cantonese contents of the respective sentences are different. The selected English words/phrases are commonly found in code-mixing speech and cover different part-of-speech categories.

The monolingual Cantonese sentences (MC) are identical to the CM sentences except that the English segments are replaced by the corresponding Cantonese words. The number of distinct MC sentences is smaller than that of CM ones because some of the English segments do not have Cantonese equivalents. Table 5 gives an example pair of CM and MC sentences.

In this example, the code-switched word “bonus” is replaced by the Cantonese word “花紅”.

Table 5. A CM sentence and the corresponding MC sentence

CM sentence:	我覺得今年有 bonus 嘅機會好渺茫。
MC sentence:	我覺得今年有 花紅 嘅機會好渺茫。
English translation:	I believe that it is very unlikely to have a bonus this year.

We also need English speech data for acoustic modeling of the English segments. Existing English databases like TIMIT and WSJ (Garofolo *et al.,* 1993; Lamel *et al.,* 1986; Paul & Baker, 1992) do not serve the purpose as they cannot reflect the phonetic and phonological properties of Cantonese-accented English. The amount of English speech data in the CM utterances is very limited. Thus, monolingual English utterances (ME) were also included as part of CUMIX to enrich the training data for the English acoustic models. The ME utterances contain English words and phrases, numbers and letters, which are most commonly used in Cantonese-English code-mixing speech.

3.2 Data Collection & Verification

The speech data in CUMIX were recorded from 34 male and 40 female native Cantonese speakers. Most of the speakers were university students. The average age was 22. The recording was carried out in a quiet room using a high-quality headset microphone. Each speaker was given a list of pre-selected sentences or phrases. He/she was requested to read each sentence fluently and naturally at a normal speaking rate. The speaker was also advised to adopt the pronunciations that they use in daily life.

Each recorded utterance was checked manually. The instants of language switching were marked. For those containing undesirable content or recording artifacts, the speakers were requested to record them again or the utterances were simply discarded. Each verified utterance is accompanied by an orthographic transcription, which is a sequence of Chinese characters with English words inserted in-between. In addition, the Cantonese pronunciations of the characters were also provided in the form of Jyut-Ping symbols.

3.3 Corpus Organization

Based on the usage, the utterances were organized into two parts, namely training data and test data. The training data set includes utterances from 20 male and 20 female speakers. Each speaker has 200 CM utterances and 100 ME utterances. Test data are intended for performance evaluation of the code-mixing speech recognition system and language boundary detection algorithms. There are 14 male and 20 female speakers in the test data. Each of them has 120 CM utterances and 90 MC utterances. Among the 34 test speakers, 5 males and 5 females were reserved as development data, which is intended for the tuning of various weighting parameters and thresholds in the system design. Table 6 gives a summary of the CUMIX corpus.

Table 6. A summary of CUMIX

		Training data	Test data
		20 male, 20 female	14 male, 20 female
CM	Duration:	7.5 hours	4.25 hours
	Duration of English segments:	1.13 hours	0.57 hours
	Total no. of utterances:	8000	3740
	No. of unique sentences:	2087	2256
	No. of unique English segments:	1047	1069
MC	Duration:		2.75 hours
	Total no. of utterances:		3060
	No. of unique sentences:		1742
ME	Duration:	1.5 hours	
	Total no. of utterances:	4000	
	No. of unique sentences:	1000	

4. Acoustic Modeling

This part of research aims at designing an appropriate phoneme inventory for acoustic modeling of Cantonese-English code-mixing speech. It is expected that some of the phoneme models are language-specific and the others are shared between Cantonese and English. Speech recognition experiments are carried out to evaluate the performances of three different sets of acoustic models in terms of syllable and word accuracy. In addition to CUMIX, two large-scale monolingual speech databases, namely TIMIT and CUSENT, are involved. CUSENT is a read-speech database developed for Cantonese LVCSR applications (Lee *et al.,* 2002). TIMIT is a phonetically balanced speech database of American English with hundreds of speakers (Garofolo *et al.,* 1993).

Table 7 explains the three sets of acoustic models, which are denoted by ML_A, ML_B, and CL, respectively. ML_A and ML_B are language-dependent phoneme models, in which Cantonese and English phonemes are separated despite the fact that some of them are phonetically similar. There are 56 Cantonese phonemes as listed in Table 8. They are adequate to compose all legitimate syllables of Cantonese. The English phoneme set has 39 elements as shown in Table 9. This phoneme set has been the most widely used in previous research (Lee & Hon, 1989). The difference between ML_A and ML_B is that they are trained with different training data as shown in Table 7.

CL is a set of cross-lingual models, designed to accommodate both Cantonese and English. As the matrix language, all Cantonese phonemes are included in the cross-lingual phoneme set. The English phonemes are divided into two parts. Phonemes that are unique to English are modeled separately, while the others are treated as Cantonese phonemes. In our work, the merging between English and Cantonese phonemes is based largely on phonetic knowledge (Chan, 2005). Due to the Cantonese accents, a number of English phonemes in the code-mixing speech are found to be sharable with Cantonese. It is also practically preferable to reduce the total number of phonemes as far as possible to facilitate effective utilization of training data. As a result, a total of 70 phonemes are selected for CL (Chan *et al.,* 2006). They are listed in Table 10. In addition to the 56 Cantonese phonemes in Table 8, a number of Cantonese diphthongs that have English equivalents are included. There are only 7 English-specific phonemes, while the others are mapped to some Cantonese equivalents.

Table 7. Different acoustic models being evaluated

Model	Phoneme inventory	Training data	
ML_A	39 English phonemes 56 Cantonese phonemes	English: Cantonese:	TIMIT CUSENT
ML_B	39 English phonemes 56 Cantonese phonemes	English: Cantonese:	CUMIX CUSENT & CUMIX
CL	70 Cross-lingual phonemes	English: Cantonese:	CUMIX CUSENT & CUMIX

Table 8. 56 Cantonese phonemes for monolingual modeling (ML_A & ML_B). Jyut-Ping symbols are used. "f-" represents a syllable-initial consonant and "-m" represents a syllable coda. "k-/kw-" means that the two initial consonants are merged as one. "s-(yu)" represents a variant of "s-" when followed by the vowel "yu".

Consonant	f-, h-, k-/kw-, g-/gw-, l-/n-, m, m-, -m, -n, ng, ng-, -ng, null, b-, p-, s-, s-(yu), z-, z-(yu), c-, c-(yu), d-, t-, w-, j-
Vowel	a, aa, o, e, eo, i, i(ng), oe, u, u(ng), yu
Vowel-stop	ap, at, ak, aap, aat, aak, ep, et, ek, ut, uk, yut, ip, it, ik, op, ot, ok, eot, oek

Table 9. English phonemes for monolingual modeling (ML_A & ML_B). APRABET symbols are used to label the phonemes.

Consonant	dh, th, f, v, w, z, zh, s, sh, t, d, b, p, ch, g, h, jh, k, l, m, n, ng, y, r
Vowel	aa, ae, ah, ao, aw, ay, eh, er, ey, ih, iy, ow, oy, uh, uw

Table 10. Phonemes for cross-lingual modeling (CL). English-specific phonemes start with the prefix "E_" and are labeled with ARPABET symbols.

Consonant (30)	f-, h-, k-/kw-, g-/gw-, l-/n-, m, m-, -m, -n, ng, ng-, -ng, null, b-, p-, s-, s-(yu), z-, z-(yu), c-, c-(yu), d-, t-, w-, j-, E_t, E_d, E_k, E_r, E_z
Vowel/diphthong (20)	a, aa, o, e, eo, i, i(ng), oe, u, u(ng), yu, iu, aai, ai, au, ou, oi, ei, E_ah, E_el
Vowel-stop (20)	ap, at, ak, aap, aat, aak, ep, et, ek, ut, uk, yut, ip, it, ik, op, ot, ok, eot, oek

The English phoneme models in ML_A are trained with TIMIT, and the Cantonese models are trained with CUSENT. The English words in TIMIT sentences are transcribed into phoneme sequences based on the CMU pronunciation dictionary (CMU). The Cantonese syllables in CUSENT utterances are transcribed into phoneme sequences using a standard Cantonese pronunciation dictionary (LSHK, 1997). All training data are assumed to follow the standard pronunciations.

For ML_B, the English phoneme models are trained with the code-switched English segments in the CM and ME utterances of CUMIX. The Cantonese phoneme models are trained with CUSENT and the Cantonese part of CUMIX. Moreover, the pronunciation dictionaries used for transcribing the utterances include not only standard English but also Cantonese-accented English and common pronunciation variants of Cantonese syllables. Thus, there may exist multiple pronunciations for a lexical entry. For each of the possible pronunciations, the acoustic likelihood of the word or syllable segment is computed. The pronunciation with the highest likelihood is adopted for the training of ML_B.

For CL, we use the same training data as for ML_B. We also use the same transcriptions as determined for ML_B except that the language-dependent phoneme symbols are converted into the cross-lingual phoneme symbols in Table 10.

The effectiveness of ML_A, ML_B, and CL are evaluated by syllable/word recognition experiments. The test data include the CM and the MC test utterances of CUMIX. The acoustic feature vector has 39 components: 13 MFCC and their first and second-order time derivatives. All phoneme models are context-dependent triphone HMMs. Each model consists of three emitting states, each of which is represented by a mixture of Gaussian density functions. States in models are clustered and tied using a decision-tree based technique with pre-set phonetic questions. ML_A and ML_B use 16 Gaussian components per state, while CL has 32 Gaussian components. The grammar network used for recognizing CM utterances is illustrated in Figure 2. For MC utterances, the recognition network is simplified into a syllable loop.

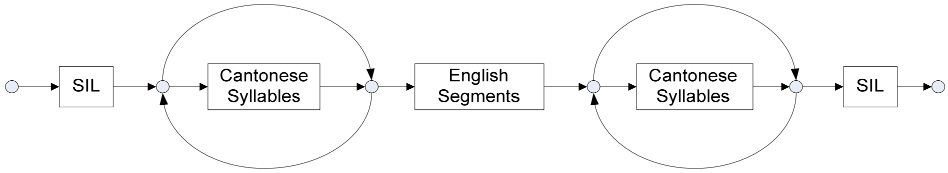


Figure 2. Grammar network for syllable/word recognition of code-mixing speech

The recognition performance is measured in terms of syllable accuracy for Cantonese and word accuracy for English. The test results are given in Figure 3. For code-switched English words, ML_A attains a very low accuracy of 18.9%. This confirms that Cantonese-accented English is very different from the native American English found in TIMIT. ML_B improves greatly in recognizing English words due to better matched training. Nevertheless, the accuracy of 40.5% is still on the low side because of the limited amount of training data and the language-dependent nature of the models. The English words in CUMIX carry Cantonese accents, such that some of the English phoneme models are very close to certain Cantonese phoneme models. In other words, similar acoustic features are captured by two different models. Hence, the confusion between English words and Cantonese syllables tends to increase. The Cantonese syllables are easily misrecognized as English words, and vice-versa. This also explains why the performance of ML_B in recognizing Cantonese syllables declines.

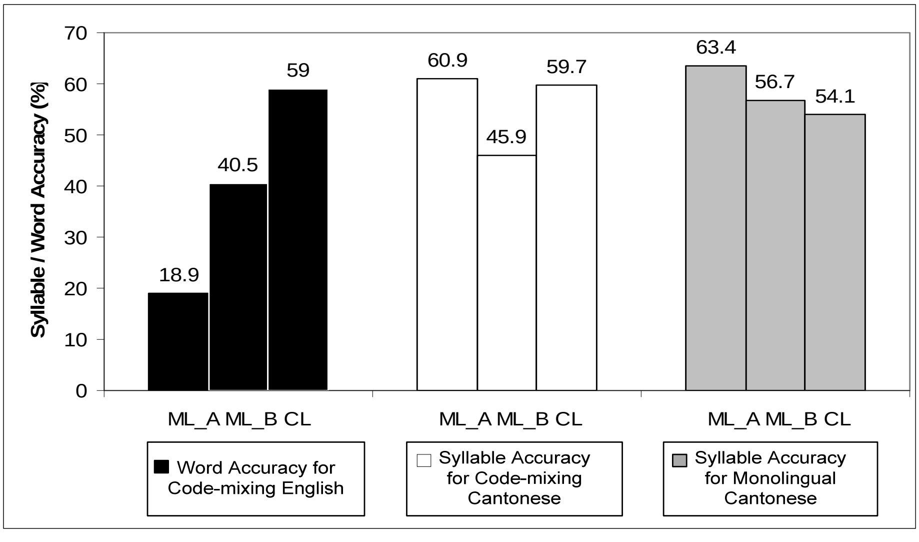


Figure 3. Syllable/word accuracy of the three acoustic models

For Cantonese speech in the code-mixing utterances (CM), the recognition accuracies attained by ML_A and ML_B are 60.9% and 45.9% respectively. The poor performance of ML_B is related to the use of language-dependent models as discussed above. The performance difference between ML_A and ML_B for monolingual Cantonese utterances (MC) is not as significant as for the CM utterances. This is because the grammar network used for MC utterances does not include an English segment, and therefore there should be no recognition error caused by the confusion between similar Cantonese and English phonemes.

CL uses a large number of shared phoneme models between English and Cantonese. It attains the best recognition accuracy of 59% for the embedded English words, and at the same time, it maintains a reasonable performance on Cantonese. It is believed that the existing design of cross-lingual models can be improved further with more understanding about the phonetic variation in code-mixing speech. More training data will also be helpful.

5. Language Modeling

5.1 Collection and Selection of Text Data

There are practical difficulties in collecting a large amount of text material to facilitate statistical language modeling for Cantonese-English code-mixing speech. Cantonese is a spoken dialect; many colloquial Cantonese words do not have a standard written form. In addition, written Cantonese is neither taught in schools nor recommended for official and documentary usage. Nevertheless, a limited amount of Cantonese text data can be found in certain columns of local newspapers, magazines, advertisements, and online articles (Snow, 2004). On the other hand, code-mixing is a domain-specific phenomenon. It is found in the discourses that involve contemporary and cross-cultural issues, *e.g.*, computer, business, fashion, food, and showbiz (Li, 1996). In our study, Cantonese text data are selected from three major sources, namely newspaper, magazines, and online diaries. Preliminary manual inspection was done to identify the sections or columns that are highly likely to contain code-mixing text. A total of 28 Chinese characters that are frequently used in spoken Cantonese but rarely used in standard Chinese were identified, *e.g.*, 嘅, 嘢, 咁 (Snow, 2004). Articles that contain these characters were considered to be written in Cantonese. As a result, a text database with 6.8 million characters was compiled. There are about 4600 distinct Chinese characters and 4200 distinct English segments in the database. About 10% of these English segments are included in the CUMIX utterances.

5.2 Training of Language Models

The text data were used to train character tri-grams. Four different models were trained:

CAN_LM: mono-lingual Cantonese language model;

CM_LM: code-mixing language model;

CLASS_LM: class-based language model;

TRANS_LM: translation-based language model.

For CAN_LM, all English words were removed from the training text. They were considered as out-of-vocabulary (OOV) words during the evaluation. OOV words are assigned zero probability so that they may be missed in recognition. For CM_LM, all code-switched segments in the training text were mapped to the same word ID during the training process, no matter whether the words were found in the training text or not. By doing so, the likelihood of English segments is made much higher than that of the Cantonese characters, thus, Cantonese words may be easily misrecognized as English words. In CLASS_LM, code-switched segments were divided into 15 classes according to their parts of speech (POS) or meanings. Most of the classes were for nouns. TRANS_LM involves English-to-Cantonese translation, by which code-switched segments are translated into their Cantonese equivalents. Nevertheless, since not all of the code-switched terms have Cantonese equivalents, the POS classes being used in CLASS_LM were considered as well.

The language models were evaluated in the phonetic-to-text (PTT) conversion task. Assuming that the true phonetic transcription is known, language models were used to determine the word sequence that best matched the transcription. For Chinese languages, PTT conversion is often formulated as a problem of syllable-to-character or Pinyin-to-text conversion. Statistical language models have proven to be very effective (Gao *et al.,* 2002). In our study, PTT conversion was treated as a sub-task of decoding for speech recognition. The proposed code-mixing speech recognition system employs a two-pass decoding algorithm (see Section 7 for details). The first pass generates a syllable/word lattice using acoustic models and bilingual dictionary. Language models are used in the second pass to decode the Chinese character sequence. PTT conversion can be done by skipping the first pass and using the true syllable-level transcription to replace the hypothesized syllable lattice. In this way, the effectiveness of language models can be assessed. The true syllable transcription of the CM test utterances is used as the input. The PTT conversion accuracy attained by different language models is given in Table 11.

Table 11. Phonetic-to-text conversion rate by different language models

Language model	PTT conversion rate (character accuracy)
Monolingual Cantonese (CAN_LM)	88.8%
Code-mixing (CS_LM)	89.3%
Class-based (CLASS_LM)	91.5%
Translation-based (TRANS_LM)	86.1%

The four language models are close to each other in performance because their differences are mainly on the code-switched segments. The translation approach (TRANS_LM) achieves the lowest PTT conversion rate. This is due to some of the translated Cantonese characters not appearing in the character list of the original Cantonese language models. This leads to the n-gram probabilities that are related to these characters being very low in TRANS_LM. The low likelihood affects the decision on the neighboring characters and leads to degradation of the overall conversion rate. Moreover, the code-switched segments are translated into Cantonese, and each translated term may contain more than one character. This causes a discrepancy in the computed values of the PTT conversion rate.

6. Language Boundary Detection

Language identification (LID) is an important process in a multilingual speech recognition system (Ma *et al.,* 2007). The language identity information allows the use of two monolingual recognizers. However, the LID for recognizing code-mixing speech is not straightforward mainly because the speech segments that can be used for decisions are relatively short. For code-mixing speech, LID can be considered as a problem of language boundary detection (LBD). We consider two approaches below (Chan *et al.,* 2006).

6.1 LBD based on syllable bigram

The syllable bigram probability of Cantonese is defined as the probability that a specific syllable pair occurs. In our study, these probabilities were computed from a transcribed Cantonese text database. In a code-mixing utterance, the Cantonese part is expected to have high syllable bigram probability, while the embedded English segments have relatively low syllable bigram probability, because of the mismatch in phonological and lexical properties. We use a Cantonese syllable recognizer based on the cross-lingual acoustic model CL as described in Section 4. For each pair of adjacent syllables in the recognized syllable sequence, the syllable bigram is retrieved. If the probability is higher than a threshold, this syllable-pair segment is considered to be Cantonese; otherwise, it is English or at the code-mixing boundary. It is possible that more than one English segment is detected within an utterance. Under the assumption that each utterance consists of exactly one English segment, we need to select one of the hypotheses. Our current strategy is to select the segment with the longest duration. On the other hand, if no English segment is found, the threshold is increased until the English segment includes at least one syllable.

To evaluate the performance of an LBD algorithm, the detected boundaries of a language segment are compared to the true boundaries. If the detection errors on both sides of the segment exceed a threshold, an LBD error is recorded. In this study, the threshold was set to 0.3 second. With the syllable bigram based detection algorithm, an LBD accuracy of 65.9%

was attained for the CM test utterances.

6.2 LBD based on Syllable Lattice

This approach makes use of the syllable/word lattice generated by a bilingual speech recognizer, which will be described in the next section. Syllable lattice is a compact representation of recognition output, which covers not only the best syllable sequence but also other possible alternatives. The lattice produced by our system contains Cantonese syllable units and English word/phrase units. English words/phrases generally have longer duration than Cantonese syllables since they may contain multiple syllables. The English segment with the longest duration in the lattice is most likely to indicate a correct recognition result, and the start and end time of the segment are taken as the language boundaries. With a properly selected insertion penalty, the LBD accuracy for CM test utterances was 82.3%.

7. A Code-mixing Speech Recognition System

7.1 System Overview and Decoding Algorithm

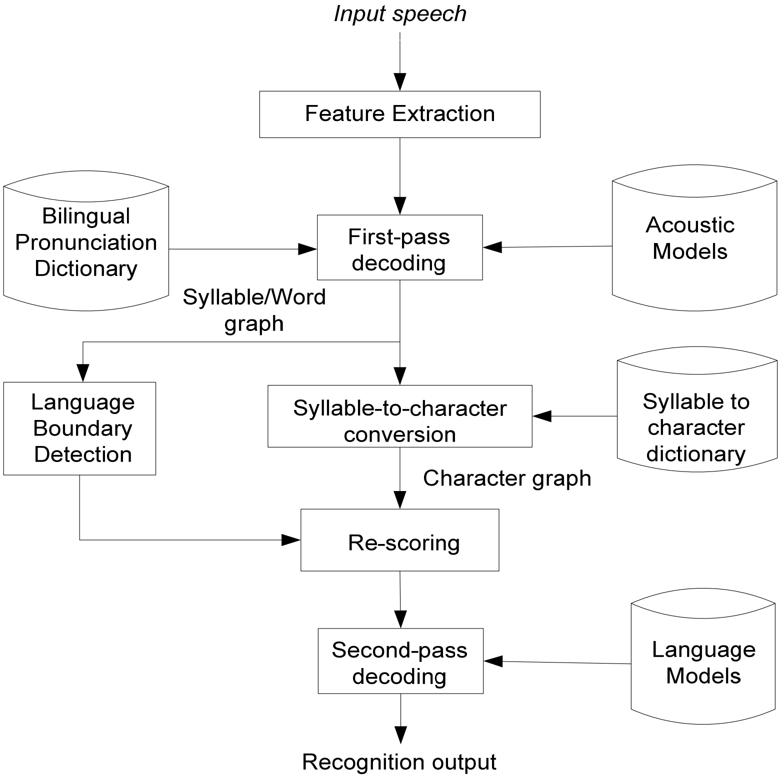


Figure 4. The proposed code-mixing speech recognition system

A code-mixing speech recognition system was developed as shown in Figure 4. It consists of the cross-lingual acoustic models, the bilingual pronunciation dictionary, and the class-based language models as described in previous sections. It is assumed that the input utterance is either code-mixing speech with exactly one English segment, or monolingual Cantonese speech. The decoding algorithm is implemented with the HTK Toolkits (Young *et al.,* 2001). It consists of two passes as described below.

First pass

In the first pass, the cross-lingual acoustic models and the bilingual pronunciation dictionary are used to construct a recognition network as shown in Figure 2. In the case where the input utterance is monolingual Cantonese, the recognition network is simplified into a syllable loop. Language models are not involved at this stage. The recognition network represents all possible hypotheses, from which the most likely ones are to be determined. The first-pass decoding is based on a token-passing algorithm. Each token refers to a partial hypothesis starting from the first frame of the utterance. At each time step, a feature vector is taken up and the existing tokens are extended through the HMM states in the recognition network. If there are many competing tokens at a network node, only the best N tokens are kept and the others are discarded. In this way, a syllable/word graph is generated as a compact representation of multiple hypotheses. The basic elements of the graph are nodes and arcs. Each arc represents a hypothesized Cantonese syllable or a hypothesized English word/phrase. It records the acoustic likelihood, the start time, and end time of the syllable or words/phrases. An example of mixed syllable/word graphs is shown in Figure 5.

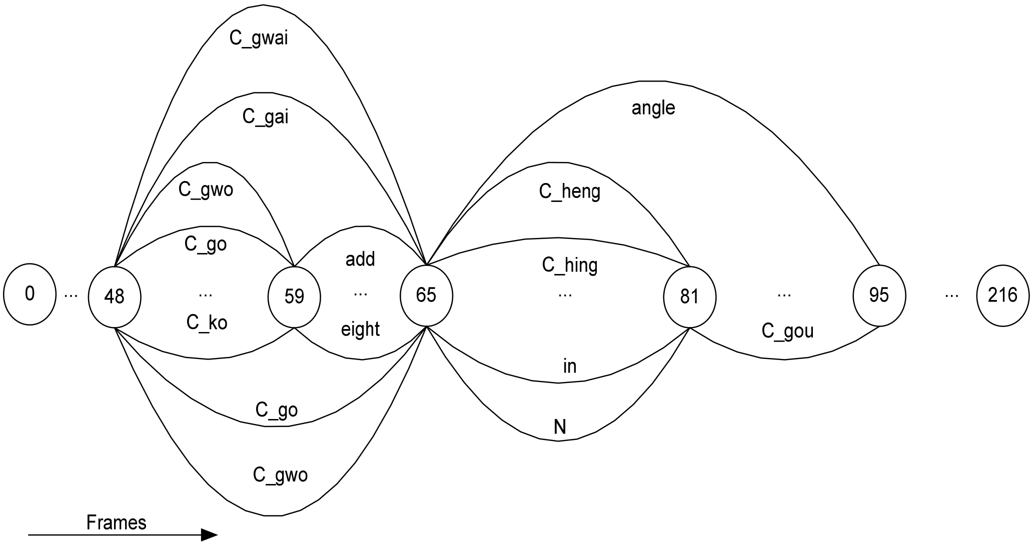


Figure 5. An example of mixed syllable/word graphs

Second pass

In the second pass, the most likely code-mixing sentence is determined from the syllable/word graph. In addition to the acoustic likelihoods, language boundary information and language models are utilized in the search process. Firstly, the language boundary information is integrated to the syllable/word lattice by modifying the acoustic likelihood of hypothesized words. If a hypothesized word is in the same language as the recognized language, the acoustic likelihood is increased by a pre-determined value; otherwise, it is decreased by the same value. The optimal value of this bonus/penalty score is derived from development data. Secondly, the modified acoustic scores are integrated with the language model scores to form a character lattice. The hypothesized syllables in the graph are mapped to Chinese characters using a pronunciation dictionary (LSHK, 1997). Since a Cantonese syllable may correspond to more than one Chinese character, the resulting character graph is in fact an expanded version of the syllable graph. The English words/phrases in the graph remain untouched. In the word graph, the posterior probability of a hypothesized word can be computed by summing the posterior probabilities of all sentence hypotheses that share the word segment w at the same time interval. In Soong *et al.* (2004), the generalized word posterior probability (GWPP) was formulated mainly to deal with the inconsistent dynamic ranges of acoustic models and language models, and with the alignment ambiguities between different sentence hypotheses. The effectiveness of GWPP has been demonstrated in Cantonese large-vocabulary continuous speech recognition (Qian *et al.,* 2006).

Let w denote a hypothesized word or syllable in the graph, with the start time s and end time t. The GWPP of w during the time interval $[s,t]$ is calculated from all word strings that contain w with a time interval overlapping with $[s,t]$, *i.e.*,

$$P([w;s,t] \,|\, x_1^T) = \sum_{\substack{W^M,\forall M \\ \exists n, 1\le n\le M, \text{ s.t.} \\ w_n=w, \text{ and} \\ [s_n,t_n]\cap[s,t]\ne\Phi}} \frac{\prod_{\mathrm{m=1}}^{M} P^{\alpha}(x_{s_m}^{t_m} \,|\, w_m)\cdot P^{\beta}(w_m \,|\, w_1^{m-1})}{P(x_1^T)} \quad , \tag{1}$$

where $W^M = \{[w_1;s_1,t_1],[w_2;s_2,t_2],\ldots,[w_M;s_M,t_M]\}$ denotes a specific word string that contains M words, and $[w_{n;}s_n,t_n]$ refers to the *n*th word in the string, which starts at time s_n and ends at t_n. The conditions of $w_n = w$ and $[s,t]\cup[s_n,t_n]\ne\Phi$ mean that the hypothesized word appears in this word string over approximately the same time interval. $P(x_{s_m}^{t_m} \,|\, w_m)$ and $P(w_m \,|\, w_1^{m-1})$ denote respectively the acoustic model scores and the language model scores. The prior probability $P(x_1^T)$ can be calculated by summing up all forward strings probabilities or backward string probabilities in the word graph. The weighting factors α and β are jointly optimized by using a held-out set of development data with a goal to achieve the minimum word error rate.

7.2 Experimental Results

The performance of the code-mixing speech recognition system in Figure 4 was evaluated using the CM and MC test utterances. For the CM utterances, the character accuracy was measured for the Cantonese part and the word accuracy is measured for the embedded English segments. From the development data in CUMIX (see Section 3.3), the best values α and β were found to be 0.009 and 1.1 respectively. This leads to an overall accuracy of 55.1% for the development utterances.

Without the use of language boundary detection, the overall recognition accuracy for CM and MC utterances were 55.3% and 50.3%, respectively, when the class-based language models CLASS_LM were used. The detailed results are given in Table 12.

Table 12. Recognition accuracy without using language boundary information

	Overall accuracy	Cantonese Character accuracy	English Word accuracy
CM test utterances	55.3%	56.0%	48.4%
MC utterances	50.3%	50.3%	

We also attempted to incorporate the detected language boundaries into the recognition process. Table 13 compares the effectiveness of the two LBD approaches described in Section 6. With LBD based on syllable bigram, the overall recognition accuracy increases from 55.3% to 57.0%. For the syllable-lattice based LBD, although the overall accuracy does not increase significantly, there is a noticeable improvement on the recognition accuracy for the English words. Among the recognition errors on English words, 39.0% of them are deletion errors, while 44.2% are substitution errors. Deletion error means that no English word is found in the top-best hypothesis string. Substitution errors are mainly caused by incorrect language boundary thus the hypothesis English word and the reference English word have no or just very little overlap in time duration. For example, the word “evening” is mistakenly recognized as “even”, and “around” became “round”.

It was also noted that the English word accuracy could be improved to 81.1% if the true language boundaries are used in the recognition process. It is believed that the recognition performance can be improved, when better language boundary detection algorithms become available.

Table 13. Recognition accuracy attained with the incorporation of language boundary information. Only CM test utterances are used.

	Overall accuracy	Cantonese Character accuracy	English Word accuracy
Without LBD	55.3%	56.0%	48.4%
LBD based on syllable bigram	57.0%	57.6%	49.0%
LBD based on syllable lattice	56.0%	56.4%	53.0%

For Cantonese, the character accuracy was close to our expectation. The character accuracy (56.4%) was roughly equal to the syllable accuracy (59.7%) multiplied by the PTT conversion rate (91.5%).

8. Conclusion

Code-mixing speech recognition is a challenging problem. The difficulties are two-fold. Firstly, we have little understanding about this highly dynamic language phenomenon. Our study clearly reveals that code-mixing is not a simple insertion of one language into another. It comes with a lot of phonological, lexical, and grammatical variation with respect to monolingual speech spoken by native speakers. Unlike in monolingual speech recognition research, there are very few linguistic studies that can be consulted. We have to understand the problems by actually working on them. Secondly, it is practically difficult to collect sufficient code-mixing data for effective acoustic modeling and language modeling. The existing CUMIX database needs to be enhanced, especially in the amount of English speech.

We have shown that cross-lingual acoustic models are more appropriate than language-dependent models. The proposed cross-lingual models attain an overall recognition accuracy of nearly 60% for code-mixing utterances. To design a cross-lingual phoneme set, we need to measure the similarity between the phonemes of the two languages. Our current approach is based on phonetic knowledge. It can be improved further with comprehensive acoustic analysis of real speech data. For language modeling, grouping English words into classes seems to be inevitable due to data sparseness. The class-based language models were shown to be effective in code-mixing speech recognition.

Two different methods of language boundary detection have been evaluated. LBD based on syllable bigram exploits the phonological and lexical differences between Cantonese and English. LBD based on syllable lattice makes use of the intermediate result of speech recognition, which is more informative than the prior linguistic knowledge. Therefore, this method attains a significantly higher accuracy than the former one in language boundary detection.

A complete speech recognition system for Cantonese-English code-mixing speech has been developed. The two-pass search algorithm enables flexible integration of additional knowledge sources. The overall recognition accuracy for Cantonese syllables and English words in code-mixing utterances is 56.0%.

Acknowledgements

This work is partially supported by a RGC grant (No. 411306) and also a grant awarded by the Shun Hing Institute of Advanced Engineering.

References

Auer, P. (1998). *Code-Switching in Conversation: Language, Interaction and Identity*. Routledge, London.

Chan, H. S. (1992). *Code-mixing in Hong Kong Cantonese-English Bilinguals: Constraints and Processes*. M.A. Thesis, The Chinese University of Hong Kong, Hong Kong.

Chan, J. Y. C., Ching, P. C., Lee, T. and Meng, H. (2004). Detection of language boundary in code-switching utterances by bi-phone probabilities. In *Proceeding of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Hong Kong, 293-296.

Chan, J. Y. C. (2005). *Automatic Speech Recognition of Cantonese-English Code-Mixing Utterances*, M.Phil. Thesis, the Chinese University of Hong Kong, Hong Kong.

Chan, J. Y. C., Ching, P. C. and Lee, T. (2005). Development of a Cantonese-English code-mixing speech corpus. In *Proceeding of Eurospeech*. Lisbon, Portugal, 1533-1536.

Chan, J. Y. C., Ching, P. C., Lee, T. and Cao, H. (2006). Automatic speech recognition of Cantonese-English code-mixing utterances. In *Proceeding of Interspeech (ICSLP)*. PA, USA, 113-116.

Chan, C.-M. (2004). Two types of code-switching in Taiwan. In *Proceeding of the 15th Sociolinguistics Symposium*. Newcastle, UK.

Ching, P. C., Lee, T., Lo, W. K. and Meng, H. (2006). Cantonese speech recognition and synthesis, In *Advances in Chinese Spoken Language Processing*, Lee, C.-H., Li, H., Lee, L.-S., Wang, R.-H., and Huo, Q., Eds. World Scientific Publishing, Singapore, 365-386.

Ching, P. C., Lee, T. and Zee, E. (1994). From phonology and acoustic properties to automatic recognition of Cantonese. In *Proceeding of International Symposium on Speech, Image Processing and Neural Networks*. Vol. 1. Hong Kong, 127-132.

Carnegie Mellon University (CMU). *The CMU Pronouncing Dictionary v0.6*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Gao, J., Goodman, J., Li, M. and Lee, K.-F. (2002), Toward a unified approach to statistical language modeling for Chinese. In *ACM Trans. on Asian Language Information Processing*, 1(1), 3-33.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L. (1993). *DARRA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*. NIST Speech Disc1-1.1, NISTIR 4930.

Gauvain, J. L. and Lamel, L. (2000). Large-vocabulary continuous speech recognition: advances and applications. In *Proceeding of the IEEE*, 88(8), 1181-1200.

Gumperz, J. (1982). *Discourse Strategies*. Cambridge University Press, Cambridge, 59.

Halmari, H. (1997). *Government and Code-switching: Explaining American Finnish*. J. Benjamins, Amsterdam.

Huang, X., Acero, A. and Hon, H.-W. (2001). *Spoken Language Processing*. Prentice-Hall, New Jersey.

Kam, P. (2003). *Pronunciation Modeling for Cantonese Speech Recognition*. M.Phil. Thesis, The Chinese University of Hong Kong, Hong Kong.

Lamel, L. F., Kassel, R. H. and Seneff, S. (1986). Speech database development: design and analysis of the acoustic-phonetic corpus. In *Proceeding of DARPA Speech Recognition Workshop*. Palo Alto, 100-109.

Lee, K.-F. and Hon, H.-W. (1989). Speaker-independent phone recognition using hidden Markov models. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37(11), 1641-1648.

Lee, T., Lo, W. K., Ching, P. C. and Meng, H. (2002). Spoken language resources for Cantonese speech processing. In *Speech Communication*, 36(3-4), 327-342.

Li, D. C. S. (1996). *Issues in Bilingualism and Biculturalism: a Hong Kong Case Study*. Peter Lang Publishing, New York.

Li, D. C. S. (2000). Cantonese-English code-switching research in Hong Kong: a Y2K review. In *World Englishes*, 19(3), 305-322.

Li, P. (1996). Spoken word recognition of code-switched words by Cantonese-English bilinguals. In *Journals of Memory and Language*, 35, 757-774.

Linguistic Society of Hong Kong (1997). *Jyut Ping Characters Table*. Linguistic Society of Hong Kong Press, Hong Kong.

Lyu, D.-C., Lyu, R.-Y., Chiang, Y.-C., and Hsu, C.-N. (2006). Speech recognition on code-switching among the Chinese dialects. In *Proceeding of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. Toulouse, France, 1105-1108.

Ma, B., Li, H. and Tong, R. (2007). Spoken language recognition using ensemble classifiers. In *IEEE Trans. on Audio, Speech and Language Processing*, 15(7), 2053-2062.

Paul, D. B., and Baker, J. M. (1992). The design for the wall street journal-based CSR Corpus. In *Proceeding of Workshop on Speech and Natural Language*. New York, USA, 357-362.

Qian, Y. Soong, F. K. and Lee, T. (2005). Tone-enhanced generalized character posterior probability (GCPP) for Cantonese LVCSR," In *Proceeding of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. Toulouse, France, 133-136.

Schultz, T. and Waibel, A. (1998). Language independent and language adaptive large vocabulary speech recognition. In *Proceeding of Interspeech (ICSLP)*. Sydney, Australia, 577-580.

Schultz, T. AND Kirchhoff, K. Eds. (2006). *Multilingual Speech Processing*. Elsevier Inc..

Shoup, J. E. (1980). Phonological Aspects of Speech Recognition. In *Trends in Speech Recognition,* Lea, W. A. Ed. Prentice-Hall, New York, 125-138.

Snow, D. (2004). *Cantonese as Written Language: the Growth of a Written Chinese Vernacular*. Hong Kong University Press, Hong Kong.

Soong. F. K., Lo, W. K., Nakamura, S. (2004). Optimal acoustic and language model weights for minimizing word verification errors. In *Proceeding of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Hong Kong.

The International Phonetic Association (1999). *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

Wester, M. (2003). Syllable classification using articulatory-acoustic features. In *Proceeding of Eurospeech*. Geneva, Switzerland, 233-236.

Wong, W. Y. (2004). Syllable fusion and speech rate in Hong Kong Cantonese. In *Proceeding of Speech Prosody*. Nara, Japan, 255-258.

Wu, C.-H., Chiu, Y.-H., Shia, C.-J. and Lin, C.-Y. (2006). Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs. In *IEEE Transactions on Speech and Audio Processing*, 14, 266-276.

You, S.-R., Chien, S.-C., Hsu, C.-H., Chen, K.-S., Tu, J.-J., Lin, J.-S. and Chang, S.-C. (2004). Chinese-English mixed-lingual keyword spotting. In *Proceeding of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Hong Kong, 237-240.

Young, S. et al. (2001). *The HTK Book (for HTK Version 3.1)*. Cambridge University, Cambridge.

Computational Linguistics and Chinese Language Processing
Vol. 14, No. 3, September 2009, pp. 305-340

Corpus, Lexicon, and Construction: A Quantitative Corpus Approach to Mandarin Possessive Construction[1]

Cheng-Hsien Chen[*]

Abstract

Taking Mandarin Possessive Construction (MPC) as an example, the present study investigates the relation between lexicon and constructional schemas in a quantitative corpus linguistic approach. We argue that the wide use of raw frequency distribution in traditional corpus linguistic studies may undermine the validity of the results and reduce the possibility for interdisciplinary communication. Furthermore, several methodological issues in traditional corpus linguistics are discussed. To mitigate the impact of these issues, we utilize phylogenic hierarchical clustering to identify semantic classes of the possessor NPs, thereby reducing the subjectivity in categorization that most traditional corpus linguistic studies suffer from. It is hoped that our rigorous endeavor in methodology may have far-reaching implications for theory in usage-based approaches to language and cognition.

Keywords: Discourse-functional Grammar, Construction Grammar, Quantitative Corpus Linguistics, Possession, Clustering.

1. Introduction

It has been observed that grammatical structures or patterns often serve as routinized formats, fulfilling specific communicative purposes in our daily interaction (Biq, 2001; Chui, 2000; Huang, 2003; Ono & Thompson, 1996; Tao & Thompson, 1994; Thompson & Couper-Kuhlen, 2005; Thompson & Hopper, 2001; Wray, 2002). Speakers' knowledge of their native languages is argued to consist of "a structured inventory of conventional linguistic units, a unit

[1] An earlier version of this paper was presented at The 2008 International Conference on Language, Communication and Cognition, Brighton, UK, 4-7 August. The author is grateful to the audience for fruitful discussion and insightful comments. Sincere gratitude also goes to Stefan Gries, Shuanfan Huang, Chiarung Lu, Iwen Su, Shuchuan Tseng for invaluable advice. Any remaining errors remain the author's responsibility.

[*] TIGP-CLCLP, Academia Sinica and National Taiwan University
E-mail: alvinworks@gmail.com

being defined in processing terms as a cognitive routine" (Langacker, 1991, p.: 511: 511). In other words, language may provide indicative evidence for our cognitive understanding of the world (Croft, 2001; Fillmore & Atkins, 1992; Fillmore, Kay, & O'Connor, 1988; Grady, 1997; Lakoff, 1993; Lakoff & Johnson, 1980; Tyler & Evans, 2003).

While considerable research has been devoted to a corpus-based approach to constructional schemas, rather little attention has been paid to the methods that are used to further "interpret" the observations. The state of art is that, after surveying the behavioral patterns of a target construction, most cognitive or discourse-functional linguists may still resort to an introspective and intuitive method to identify its sub-patterns. While we do not wish to deny the important role that introspection ultimately plays in the advancement of theorizing, we expect a bottom-up procedure may lend more objectivity, thus credibility, to the empirical results. Therefore, a burgeoning research paradigm - corpus linguistics - now utilizes corpora to investigate the usage patterns and the semantic profiles of these conventional schemas in pursuit of a thorough understanding of our cognitive conceptualization.

A traditional corpus linguistic study on discourse-functional or cognitive grammar often adopts the following approach, as shown in Figure 1 (Biq, 2004a, 2004b, 2004c; Chang, 2002; Chui, 2000; Liu, 2002; Su, 1998, 2004; Tao, 2003b; Wang, Katz, & Chen, 2003).

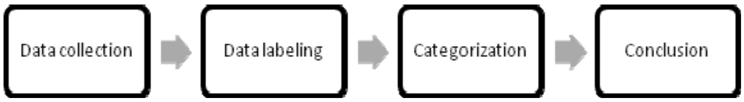


Figure 1. A typical procedure for traditional corpus linguistic studies

Initially, all the target constructions are collected from a corpus (Data collection). Second, all the relevant target constructions are manually labeled according to some researcher-defined features (Data labeling). Third, based on those manually-labeled features, the analyst tries to identify the "types" of these constructions and generate descriptive statistics to obtain a general distribution of the categories identified (Categorization). Finally, conclusions and implications are drawn on the basis of the constructional types of the highest raw frequency counts (Conclusion). It should be noted that such a working pipeline in traditional corpus linguistics has established itself in previous decades as more and more researchers submit to the view that corporal data reflect our grammatical knowledge. Therefore, if one commits him or herself to such a functional view of grammar, one would first collect data from the corpus, label them, categorize them into groups, and make generalizations based on the collected data. We take issue with the detailed procedure of how each step in the pipeline is achieved, namely, how the data is collected, how the data is labeled, and how the data is categorized.

Take Biq's (2004b) study on the patterns of Mandarin stative verb *hao* 'good' for example. In order to find the co-occurrence patterns of *hao*, she first collects all the relevant instances from a 15-hour spoken database and narrows her emphasis to two collocation patterns *hao* + *le* and *hai* + *hao*. Instances containing the target pattern are then further categorized on the basis of her operationally-defined syntactic and pragmatic criteria, and the distribution of these identified types is given in Table 1.

Table 1. The various senses/functions of hao+ le found in conversational data

SENSE/FUNCTION	NUMBER	PERCENTAGE
Topic transition	11	7.10%
Hao = resultative	13	8.30%
Hao = SV	21	13.50%
Conditional	46	29.50%
Recommendatory	65	41.60%
TOTAL	156	100.00%

Finally, conclusions are drawn based on the distribution of the type frequency. While such a traditional corpus linguistic approach has been widely adopted by most discourse-functional and cognitive grammarians, several methodological issues may merit more careful consideration.

In the first step of a traditional corpus linguistic study, the size of the corpus has long been a controversial issue in that small samples may undermine the validity of the results. For instance, even though "Recommendatory" serves as the most frequent type in Biq's observation, 156 tokens may still undermine the credibility of the distribution or even increase the possibility of the by-chance observation. Nevertheless, for most discourse-functional linguists who work on spontaneous speech, the problem of the sample size may appear inevitable due to the enormous manual labor of spoken corpus construction. Given this limitation of the spoken corpus, it is suggested that analysts might as well pursue further statistical analysis so as to increase the confidence level of their numbers. That is, given the maximal recall of the target construction in a corpus of considerable size, it would be theoretically more convincing if the distribution could be statistically tested so as to compensate for the deficiency in small-scale sampling.

With respect to the second step, the features for identifying the types of the target construction are often criticized for being researcher-dependent and lacking basis for cross-analyst comparison. In the case of Biq's study, linguists may differ in how they categorize the pragmatic functions of *hao* + *le*, thus leading to difficulty in comparing different analyst's categorizations of the same construction. Of crucial importance is the third step in a traditional corpus linguistic approach, where only the distribution of the raw frequency is consulted when conclusions are being drawn. While we acknowledge the fact that frequency plays a crucial role in the formation of our grammatical knowledge (Bybee, 2005; Bybee & Hopper, 2001), we believe that such frequency effects should exclude the by-chance possibilities due to sampling in corpus linguistics. In Table 1, chances are that in daily interaction, recommendatory speech acts may be frequent in general, thus contributing to its higher frequency in the *hao* + *le* co-occurrence patterns. If that is the case, it could be argued that, for all the constructions capable of performing recommendatory acts, this use will eventually emerge as the most frequent type among its pragmatic categorization. In other words, inferential statistics are needed to test whether the recommendatory act is indeed far more frequent than expected. In view of these potential challenges, a quantitative corpus linguistic approach has emerged (c.f. Baayen (2008) for an overview).

In a quantitative corpus linguistic study, the analyst's subjectivity is hoped to be reduced to a minimum. In terms of sampling, (semi-)automatic retrieval of the target pattern is usually adopted to ensure a better recall rate in a large balanced corpus. Second, the features for categorizing the constructional tokens are rigorously *quantified* in an operationally-defined way so that inter-analyst comparison of the results can be easily made. Most crucially, the categorization of the target patterns is made in a bottom-up procedure to replace the analyst's manual efforts as well as subjective factors. Gries and Stefanowitsch (to appear) adopts hierarchical agglomerative cluster analysis to objectively determine semantic classes of constructional sub-patterns. Furthermore, hierarchical cluster analysis has proven itself useful in a wide range of linguistic analyses such as semantic profiles of polysemy (Divjak & Gries, 2006), typology (Croft, 2008), language phylogeny (Atkinson & Gray, 2005; Dunn, Terrill, Reesink, Foley, & Levinson, 2005; McMahon & McMahon, 2003), grammaticization (Hilpert, 2007), and language development (Wiechmann, 2008). Such sophistication in the analytic process facilitates the communication between discourse-functional grammarians of different research paradigms.

The present study, therefore, aims to investigate the interaction between lexicon and construction in a quantitative corpus-based approach. With special focus on a case study of Mandarin Possessive Construction (NP1-*DE*-NP2), this paper addresses one fundamental question for every potential constructional schema: Does this constructional schema have any basic semantic patterns or any other sub-patterns? Specifically, the predictions are: 1) If NP1-*DE*-NP2 Construction has a basic meaning, the NP1-NP2 pairs will yield us such semantic sub-patterns as the major category. 2) If NP1-*DE*-NP2 Construction has *no* basic meaning, the NP1-NP2 pairs will yield us some other heterogeneous sub-patterns, or none. Meanwhile, we will compare the rank-ordering of the raw frequency counts in a traditional corpus linguistic approach with our sophisticated measures to illustrate the potential danger in relying on the former for theorizing.

The rest of the paper is structured as follows. In Section 2, a brief layout of our methodological framework - collostructional analysis - is introduced with a special emphasis on the covarying collexeme analysis. Section 3 will briefly describe the data source and our research methods and demonstrate the inferential statistics used in the evaluation of our data. Results and discussion will be provided in Section 4, illustrating the weaknesses of traditional corpus linguistic studies and the strengths of quantitative corpus linguistic studies. Section 5 concludes this paper with directions for future research and theoretical implications.

2. Lexicon and Construction

While the importance of constructional schemas has come to be the central focus of discourse-functional grammarians (c.f., Croft & Cruse, 2004), it still remains unclear how these constructional analyses can be compared and evaluated given that different linguists resort to different evidence and methods. For instance, some linguists may base their description of the constructional profiles on their own native intuition without quantitative corpus data (Fillmore, *et al.*, 1988; Kay & Fillmore, 1999; Langacker, 2003; Michaelis, 2003; Michaelis & Lambrecht, 1996; Tyler & Evans, 2003). Traditional corpus linguists may take a step further to capitalize on the raw frequency distribution of the words occurring in the target construction (Biq, 2004a; Dancygier & Sweetser, 2000; Goldberg, 1998; Liu, 2002; Su, 2002, 2004; Wang, *et al.*, 2003). Methodologically speaking, little headway has been made in examining the statistical validity of the traditional quantification and little attempt has been made to define an operational method for an analyst to generate the semantic classes of a constructional schema. Occupying the niche, collostructional analysis, proposed by Stefanowitsch and Gries (2003), provides a more rigorous approach to identifying the meaning of a grammatical construction.

Collostructional analysis represents one rigorous corpus-based methodology in discourse-functional linguistics. It makes theoretical commitments to a holistic and symbolic view of linguistic units and, at the same time, bases its quantitative methods on sophisticated statistical analyses. This empirical approach not only flavors the research of usage-based grammar with a more serious emphasis on statistical evaluation but also refreshes the direction of corpus linguistics with a more construction-specific focus on lexico-structural relations. It serves as an umbrella term, referring to research that investigates the correlation/association between words and constructional schemas.

We would like to briefly introduce the terminology and principles in collostructional analysis for the ease of the following exposition. First, lexemes that are attracted to a particular construction are referred to as *collexemes* of the construction. Crucially, collostructional analysis considers the overall distribution of the words in the corpora in calculating the association strength of those words to a specific constructional schema. The association strength between a collexeme and a construction is measured by submitting all the raw frequency counts of each word in the specific slot of the construction to the Fisher-Yates Exact Test (Pedersen, 1996). Each word occurring in the slot of the construction will be ordered by *collostrength* - defined as the log-transformed *p*-value (to the base of 10) with a positive/negative sign that indicates attraction/repulsion to the construction. This association measure allows a cognitive linguist to probe into human conceptualization through a quantitative study of the relation between words and constructional schemas.

In addition, as constructional schemas often encode a relational meaning, observations on pairs of collexemes in a construction may play an even more crucial role in the identification of the construction semantic profile. Under a usage-based cognitive-linguistic framework, grammatical patterns have been studied in terms of colligations, *i.e.*, linear co-occurrence preferences and restrictions held between specific lexical items and its surrounding syntagmatic contexts (Bybee & Scheibman, 1999; Hunston & Francis, 1999; Scheibman, 2002; Thompson, 2002; Thompson & Hopper, 2001). All of these findings point to the hypothesis that the meaning of one construction relies on the words co-occurring most often with the construction. The assumption behind this reasoning is: a word may occur in a construction if it is semantically compatible with the meaning of the construction (Goldberg, Casenhiser, & Sethuraman, 2004; Stefanowitsch & Gries, 2005). Following this hypothesis, we would expect that, given a construction with two variable slots, observations on the co-occurring patterns in these slots may yield useful empirical evidence for the (semantic) relation encoded by the construction. In this respect, Gries and Stefanowitsch (2004) extend collostructional analysis to covarying collexeme analysis, and seek pairs of collexemes that are statistically attracted to each other within a construction (*i.e.*, *covarying collexemes*). Furthermore, Gries and Stefanowitsch (to appear) further adopt a clustering-based approach to identify the potential sub-patterns of covarying collexemes in reflection of the semantic profiles of the target construction.

The present study is by and large compatible with Gries and Stefanowitsch (to appear), to which it is indebted for part of its general outlook, but poses some rather different questions, which we will identify in Section 3. Therefore, in order to investigate the semantic coherence of MPC, a closer look at the correlation between NP1 and NP2 in MPC may present itself as a rewarding endeavor. In Section 3, we will provide a more detailed illustration of our hypotheses and methods.

3. Method

The present study adopts a quantitative corpus-based approach. Initially, the data was collected from the Academia Sinica Balanced Corpus of Mandarin Chinese. This is the major Chinese corpus with detailed parts-of-speech tagging, and it includes a fairly wide range of genres and styles (mostly formal registers). Instances of Mandarin Possessive Construction (MPC) "NP1 + *DE* + NP2" were automatically retrieved via regular expressions[2]. Retrieval of the constructional instances was done in Java scripts written by the author.

Subsequently, we looked for quantified operationally-defined features to further categorize our MPC tokens. Unlike a traditional corpus linguistic approach, we aimed to reduce the involvement of the analyst's judgment to a minimum. Nevertheless, after collecting instances of the target pattern, traditional corpus linguists usually adopt two types of methods to categorize the target construction. One method is to first formulate the possible semantic categories that the target construction tokens may belong to, then label each token with an appropriate category label. In other words, this approach packages all the categorization process into the analyst's mind and the reader could only see the overall distribution of these predetermined semantic categories. How each target token is categorized into certain semantic category (*i.e.*, the operationally defined criteria) is often obscure to the readers. The other method that a traditional corpus linguistic study may adopt is to formulate a set of researcher-dependent features, usually nominal variables, then manually mark each token with the values of each feature. Then, the analyst categorizes all of the tokens according to the feature values in an introspective fashion. The disadvantage of this approach is obvious. On the one hand, the features tagged for each token usually vary from linguist to linguist and are often categorical and not quantified. On the other hand, even though the features are operationally workable, an introspective way of categorization invites a considerable degree of subjectivity in determining the clusters from the dataset. For instance, the semantic relations can be summarized into 10 labels as in Stefanowitsch (2003) or can be further elaborated into 35 as in Moldovan *et al.* (2004). Different linguists may have different labels and it would be hard to determine if two similar labels are truly semantic equivalents in both analysts' minds, thereby reducing the possibility of comparing the conclusions from different studies. Therefore, both ways of traditional corpus linguistic studies may lead to difficulty in comparing research findings with each other. More challenges to traditional corpus linguistic approach will be elaborated in Section 4.3.

Following Gries and Stefanowitsch (to appear), we adopt a specific type of hierarchical clustering algorithm known as neighbor-joining clustering. A typical process of hierarchical cluster analysis includes: 1) comparing pairwise (dis)similarities between the items in a (dis)similarity matrix via a vector-based representation of the items; 2) successively

[2] Based on the POS tagging principles elaborated in CKIP Technical Report 95-02/98-04, only nouns tagged as Na, Nc, Nd, and Nh were included as our relevant MPC constructional instances. We excluded proper names (Nb), determiners (Ne), classifiers (Nf), and postpositions (Ng). Furthermore, for all the nouns preceding *DE*, we retrieved the rightmost noun as our possessor NP1; for all the nouns following *DE*, we retrieved the first noun tagged with Na, Nc, Nd, or Nh as our possessed NP2.

amalgamating all items into clusters based on the (dis)similarity matrix, which reaches maximal intra-cluster similarity and inter-cluster dissimilarity; 3) visualizing the hierarchical structure of the datasets in the form of a tree-like dendrogram. Specifically, neighbor-joining clustering is often used in phylogeny estimation in biology (Saitou & Nei, 1987), aiming to reconstruct phylogenetic trees from evolutionary distance data under the principle of minimum evolution. Dunn *et al.* (2005) also successfully extends neighbor-joining clustering to the reconstruction of phylogeny in Oceanic languages. Our reason for choosing this algorithm lies in the assumption that constructional semantic profiles evolve similarly to phylogenic evolution in the sense that different semantic patterns of a construction, like different senses of a lexical word, may form a structured polysemy (Goldberg, 1995; Tyler & Evans, 2003). Furthermore, it is suggested that structured polysemy usually emerges from the conventional usage of high frequency via conceptual mechanisms of metaphor and metonymy (Hopper & Traugott, 1993; Traugott & Dasher, 2002; Tyler & Evans, 2003). In other words, semantics of a constructional schema is argued to evolve with language use (Bybee, 1998; Hopper, 1987; Huang, 1998; Tao, 2003a). It is this emergent or evolutionary nature of grammar and semantics that leads us to the decision of adopting phylogenetic clustering in our study. Specifically, in neighbor-joining clustering, not every node on the bottom should be collapsed into one ancestor node. This flexibility allows the possibility that not every sub-pattern comes from one prototypical pattern of the constructional instantiations.

From a perspective of the discourse-functional approach to language, the meaning of a word or a construction is defined by how speakers use it in their daily interaction (Scheibman, 2002; Tao, 2003b; Thompson & Couper-Kuhlen, 2005). In order to look for the semantic coherence encoded by MPC, two possibilities may be pursued: 1) to cluster NP1 based on NP2; 2) to cluster NP2 based on NP1. In the present study, we choose the former approach on a discourse functional basis. It has been observed that the possessor NP in MPC often serves as a topic to which new information encoded by the possessed NP is attached. Therefore, the clustering patterns of the NP1 may shed light on the overall semantic domains of the MPC instance. Furthermore, in the MPC context, the cooccurrence pattern of each NP1 with their NP2 may serve as traces on how speakers frequently make reference to the possessor NPs, thus reflecting the semantic coherence of NP1. That is, a look at how each NP1 is correlated with different types of NP2 in MPC may shed light on their similarity in their references of their possessed entities. If two types of NP1 are correlated with similar types of NP2, they are more inclined to form a semantically coherent class, where their possessed entities are of great similarity. For instance, if in NP1 position, *shi4chang3* ‘market’ and *chan3pin3* ‘product’ often co-occur with similar groups of NP2 such as *gong1zuo4* ‘job’, *xu1qiu2* ‘demand’, *qing2kuang4* ‘condition’, *fan3ying4* ‘reaction’ in MLC, they may easily form a cluster, thus suggesting their similarity in their reference of their possessed entities (*i.e.*, both being conceptualized as consisting of similar groups of entities). Also, if an abstract entity and a concrete entity are clustered together at an early stage, they may be argued to bear great resemblance in metaphorical conceptualization. Based on these correlation patterns, we can then infer if semantic coherences do exist among different types of NP1. Our working assumption is:

-If MPC has coherent meanings, the NP1 clustering will yield us such semantic sub-patterns as prominent categories at the early stage;

-If MPC has grammaticized as a pure syntactic formative, the NP1 clustering will yield us more heterogeneous sub-patterns, or none.

As clustering approaches are sensitive to the problem of data sparseness and often yield their best results when applied to moderately frequent cases (Kaufman & Rousseeuw, 2005 [1990]), we make a compromise that strikes a balance between the representativeness of the sample and the efficiency of the algorithm. Figure 2 shows the relationship between covarying NP frequency threshold and data preservation percentage. We choose to include 83% of the NP2 by setting a threshold of 5 for the frequency of NP2 and cluster only the top 100 frequent NP1, amounting to 32% of all the NP1. That is, only those covarying NP2 occurring at least 5 times in our original dataset are considered a feature for NP1 in the subsequent vector representation and clustering.

After data filtering, we transform each type of NP1 into vectors based on their association with each covarying collexeme NP2, as tabulated in Table 3. Now that we have a definition for the features or dimensions of each NP1 vector (NP2 of frequency larger than 5), we need measures of association between each NP1 and a given feature (*i.e.*, each type of NP2). It has been observed that cooccurrence raw frequency, as shown in Table 3, is a poor measure of association between a word and a context feature (Jurafsky & Martin, 2008 [2000], p.: 661: 661; Manning & Schütze, 1999, p.: 156: 156). We may require a weighting or measure of association that asks how much more often than chance the feature co-occurs with each type of NP1. Following Gries and Stefanowitsch (to appear), we adopt collostrength from covarying collexeme analysis as our measure of association between each type of NP1 and its covarying NP2 feature.[3]

[3] For further justification for the use of *p*-values as a measure of association strength, please refer to Footnote 6 in Stefanowitsch and Gries (2003). In this analysis, we also tried *t* score as our measure of association, as suggested in Manning and Schütze (1999), and the results were similar to what we had obtained from collostrength measure.

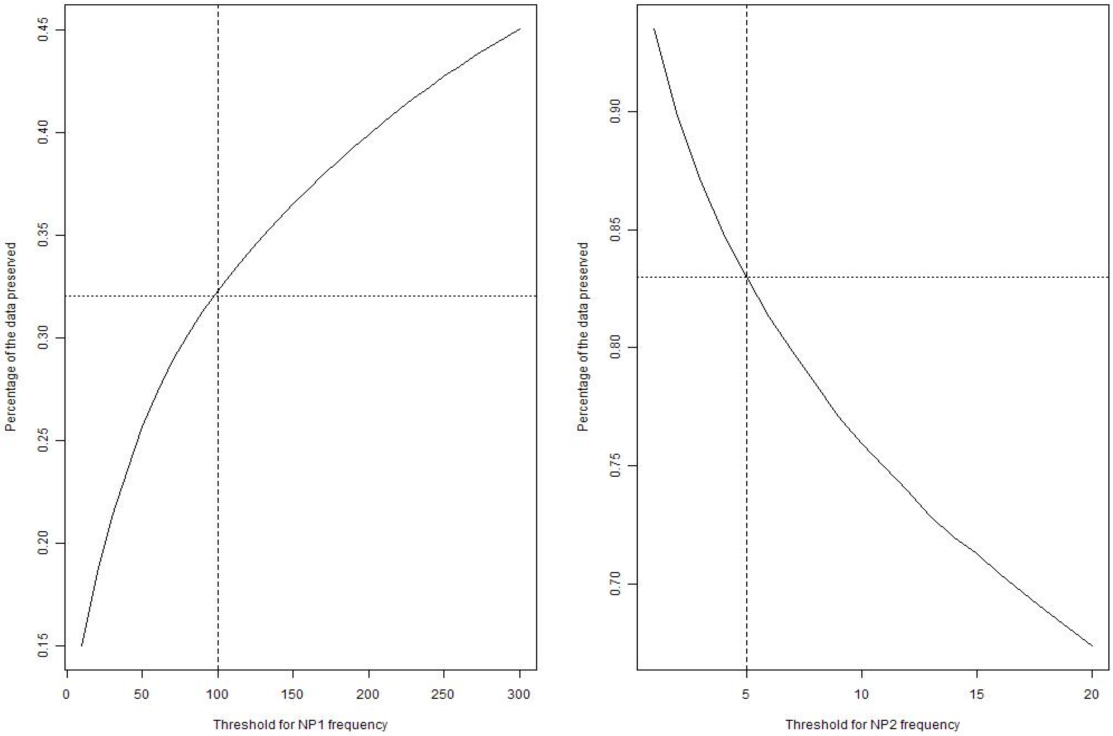


Figure 2. Threshold for NP1 and NP2 and the percentage of data perserved

For example, let us consider the distribution of *zheng4fu3* ‘government’ and *zheng4ce4* ‘policy’ in MPC (*i.e.*, *zheng4fu3 DE zheng4xe4* ‘the policy of the government’) as shown in

Table 2 (parentheses indicate expected frequencies and italics indicate observed frequencies). Applying the Fisher-Yates Exact test to this table yields a *p*-value of 1.11e-59, corresponding to a $p_{\log 10}$-value, *i.e.*, collostrength, of 58.95. This extreme *p*-value indicates that the association between *zheng4fu3* and *zheng4ce4* in MPC is a relatively strong one.

Table 2. The distribution of* zheng4fu3 *and* zheng4ce4 *in Mandarin Possessive Construction

	zheng4ce4	Other NP2	Row Totals
zheng4fu3	40(*1*)	410(*449*)	450
Other NP1	230(*269*)	207829(*207790*)	209059
Column Totals	270	208239	208509

Table 4 shows part of the co-occurrence table with the collostrength of each covarying collexeme pair in the cell. Higher collostrength may suggest a stronger association between

NP1 and NP2.

Table 3. Co-occurrence table of the NP1 (row) with the covarying NP2 (column) in MPC (raw frequency count as assoication measure)

NP1 NP2	*ren2* 'man'	*sheng1huo2* 'life'	*xin1* 'heart'	*wen4ti2* 'problem'	*hai2zi5* 'child'	*she4hui4* 'society'	…
ta1 'he'	63	49	49	17	23	6	
wo3 'I'	46	35	152	25	90	2	
zi4ji3 'self'	24	102	26	26	55	8	
ren2 'man'	21	44	40	20	5	6	
ta1 'she'	27	21	39	12	18	1	
wo3men5 'we'	11	55	25	8	48	110	
ni3 'you'	21	25	28	11	38	3	
ta1men5 'they'	11	45	19	8	18	11	
Tai2wan1 'Taiwan'	13	9	3	8	2	18	
…							

Table 4. Co-occurrence table of the NP1 (row) with the covarying NP2 (column) in MPC (collostrength as assoication measures)

NP1 NP2	*ren2* 'man'	*sheng1huo2* 'life'	*xin1* 'heart'	*wen4ti2* 'problem'	*hai2zi5* 'child'	*she4hui4* 'society'	…
ta1 'he'	0.248008	0.063422	1.211847	1.18E-05	0.006771	0	
wo3 'I'	0.040219	0.006473	Inf	0.02988	Inf	0	
zi4ji3 'self'	6.10E-07	Inf	0.045449	0.064862	5.841638	6.70E-07	
ren2 'man'	0.019297	3.341083	5.632644	0.548443	0.000205	0.000827	
ta1 'she'	0.171652	0.071817	5.323306	0.033664	0.550925	8.00E-08	
wo3men5 'we'	3.46E-05	7.69897	1.44626	0.003065	Inf	Inf	
ni3 'you'	0.187166	0.858615	3.448672	0.110468	Inf	0.00024	
ta1men5 'they'	0.003388	7.522879	1.313479	0.037517	1.585381	0.298076	
Tai2wan1 'Taiwan'	0.03479	0.008238	0.000383	0.07051	0.000245	2.061541	
…							

Next, we compute pairwise distance matrix among these 100 types of NP1. As summarized in Jurafsky and Martin (2008 [2000]: 663-667), correlation similarity measures are more prone to detect and to use curvature of vectors in multidimensional space; these measures may work better for word similarity in information/document retrieval as compared to distance dissimilarity measures. Moreover, according to Manning and Schütze (1999: 299), among all the distance-based measures, the cosine is the most frequently-used measure in the comparison of semantic similarity (*c.f.*, Curran (2004)).[4] Therefore, we compute a pairwise cosine distance matrix and submit this matrix to neighbor-joining clustering. The statistical evaluation is computed in R scripts written by the author, using the *ape* package developed by Paradis (2004).

Furthermore, we compare the lists ordered by raw frequency and collostrength, respectively, by submitting these two rank-orderings to Friedman's rank test for correlated samples. This test is the nonparametric analogue of the one-way repeated-measures ANOVA, often being applied to test if two rank-orderings differ significantly. By so doing, we demonstrate the degree to which raw frequency overlaps with the collostrength, thus highlighting the potential danger in relying on the raw frequency in theorizing.

Before the discussion of the results, let us briefly turn to the question of why we chose Mandarin Possessive Construction as our pilot study. Even though we name this construction as "possessive" here, its constructional meaning is not as uncontroversial as the naming suggests. The reason for choosing this as our target construction is mainly due to the cross-linguistic complexity of possessive or genitive constructions (Baron, Herslund, & Sorensen, 2001; Dong, 2003; Heine, 2001; Lyons, 1986; Nikiforidou, 1991; Stefanowitsch, 2003; Taylor, 1996). As implied in its alias as "associative phrases" (Li & Thompson, 1981), MPC has been notorious for its encoding of diversified semantic relations between two NPs to the extent that Li and Thompson (1981) even argue that "the precise meaning…is determined entirely by the meanings of the two noun phrases involved". While a typical possessive construction may encode a semantic relation of "possession", including ownership, kinship, and component-part relations (Nikiforidou, 1991; Stefanowitsch, 2003; Taylor, 1996), it is still unclear whether MPC indeed exhibits semantic coherences in its distributional patterns, or should better be analyzed as a semantic-general syntactic formative. Hopefully, the empirical evidence from the covarying collexemes may help solve this controversial issue.

[4] Curran (2004) evaluated a wide range of similarity measures by comparing the results with gold-standard thesauri and concluded that Dice and Jaccard methods perform best as measures of vector similarity. As a result, we also computed the similarity matrix based on these methods and submitted them to hierarchical clustering. The results were by and large similar to what we had obtained from the cosine similarity measure. Therefore, we shall base our discussion on the results from the cosine similarity measure.

4. Results and Discussion

208509 tokens of MPC were extracted from the Academia Sinica Corpus of Mandarin Chinese. These MPC instances consist of 26005 types of NP1 and 25987 types of NP2, amounting to 159645 types of NP1-NP2 pairs, *i.e.*, covarying collexemes. Each distinct type of NP1-NP2 was submitted to statistical evaluations, and the results are as follows.

4.1 Semantic Coherence in MPC

We cluster NP1 according to its covarying NP2 in MPC. After filtering our infrequent NP1 and NP2 types, we cluster the most frequent 100 NP1 according to their covarying NP2 of frequency larger than 5. This boils down to a 100 by 4372 contingency table with 58470 tokens of MPC in total, as shown earlier in Table 3. All of the possessor NPs (NP1) are then transformed into vector representations based on their collostrength with each type of covarying collexeme NP2, as shown previously in Table 4. The cooccurrence measures between NP1 and NP2 serve as criteria for classification of the NP1. We then compute the cosine distance between each pair of NP1 and submit the distance matrix to neighbor-joining clustering to obtain a tree-like representation of the NP1 categorization.

In a tree size with 100 tips (*i.e.*, 100 types of NP1), the information that is supposed to be summarized is likely to be no longer visible. Therefore, instead of plotting out the whole tree, obscuring the clustering information that is sought, we choose to plot only a portion of the full dendrogram at a time, while indicating its context - how it relates to the rest of the tree. In the following illustration, the original dendrogram is divided into three parts, where the whole tree is plotted on the left and the subtree on the right. The location of the subtree is indicated with the color on the whole tree.

The results from Figure 3 to Figure 5 are moderately revealing in that several coherent semantic frames in NP1 emerge from the dendrograms. Specifically, 7 semantically coherent categories emerge from the amalgamative process: Human, Time, Country, Enterprise, Culture, Knowledge, and Institution. Based on these correlation patterns, we suggest that semantic coherences do exist among different types of NP1, supporting the claim that MPC has not fully grammaticized as a pure syntactic formative.

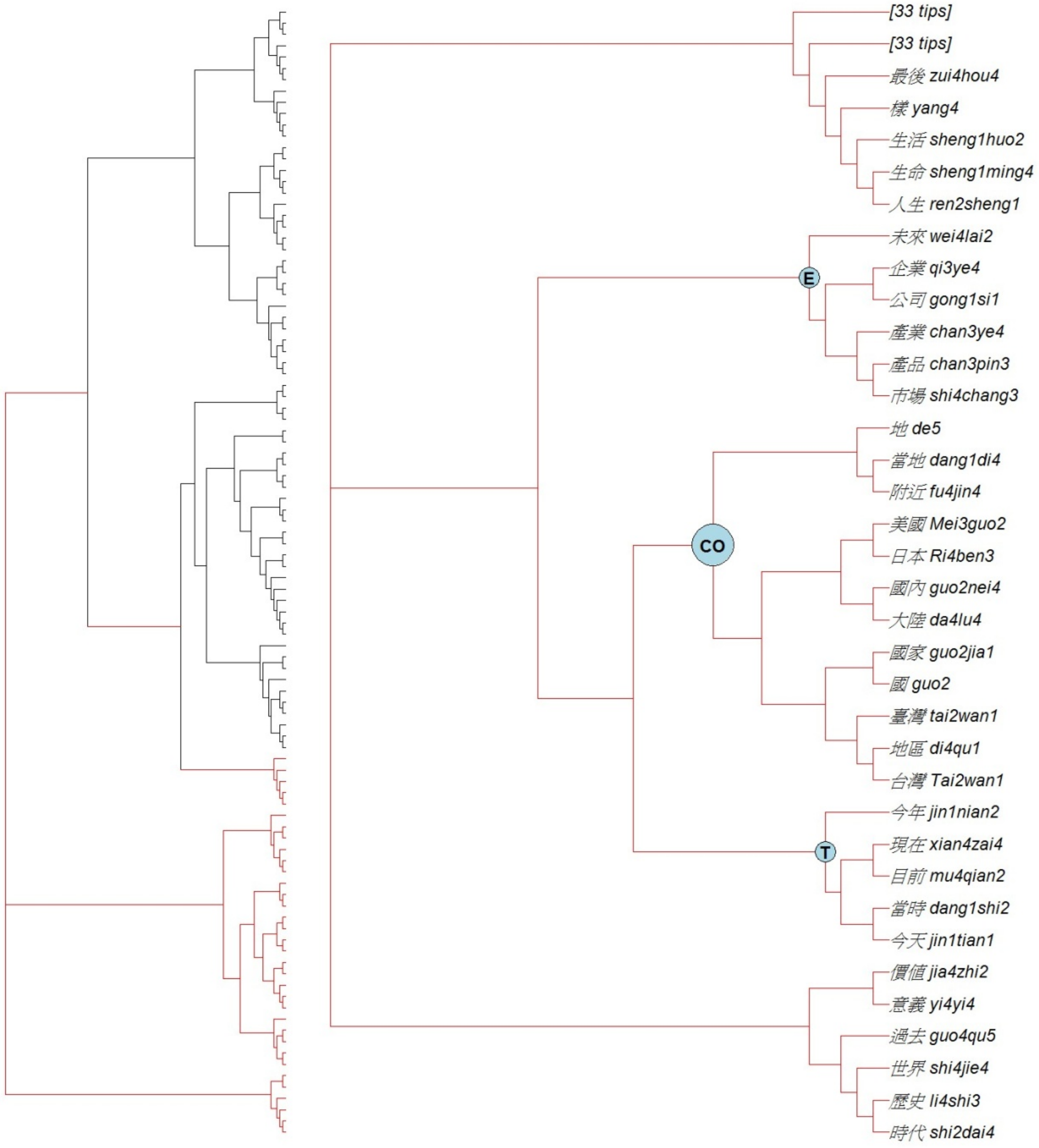


Figure 3. Subtree one of the dendrogram

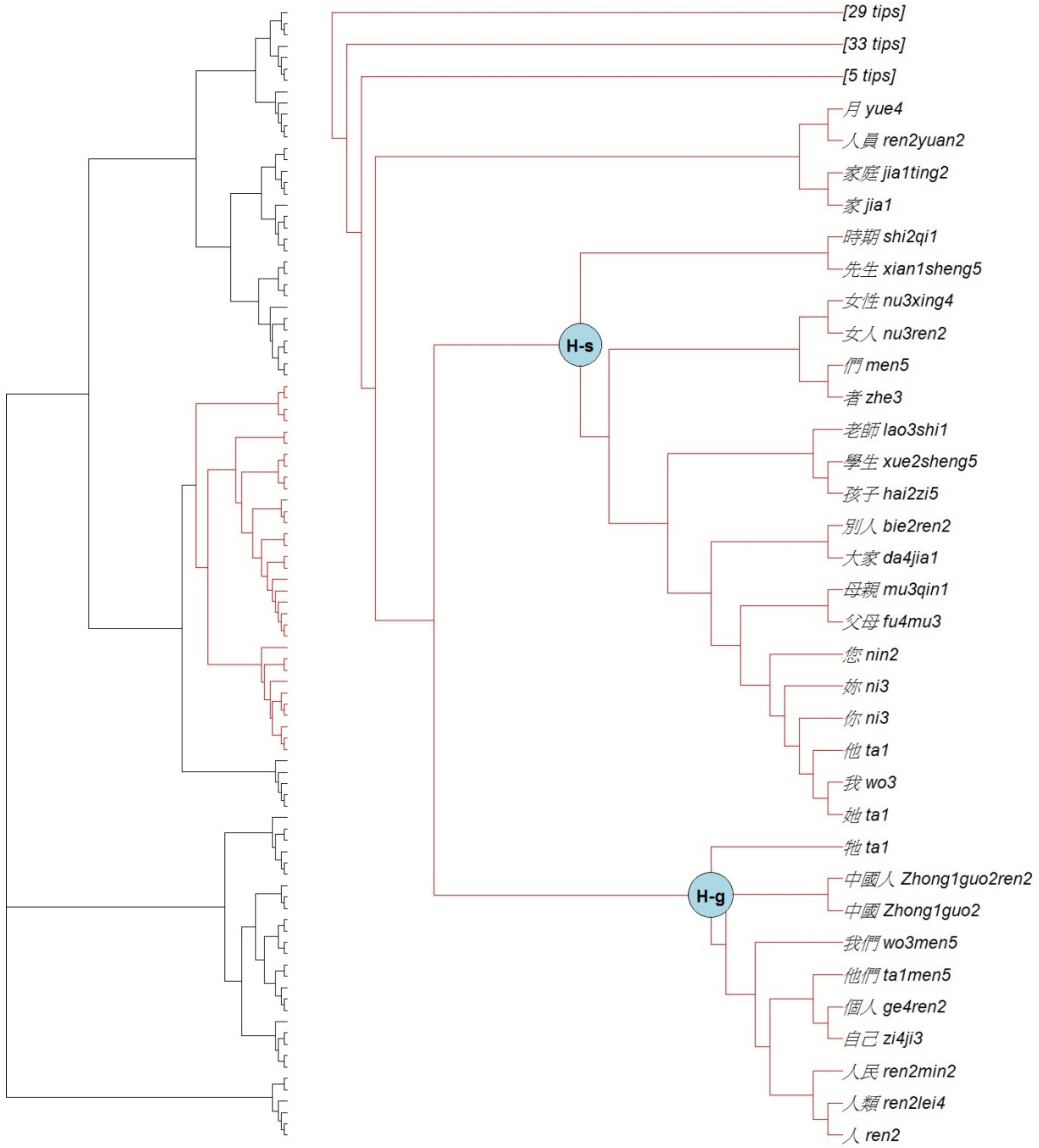


Figure 4. Subtree two of the dendrogram

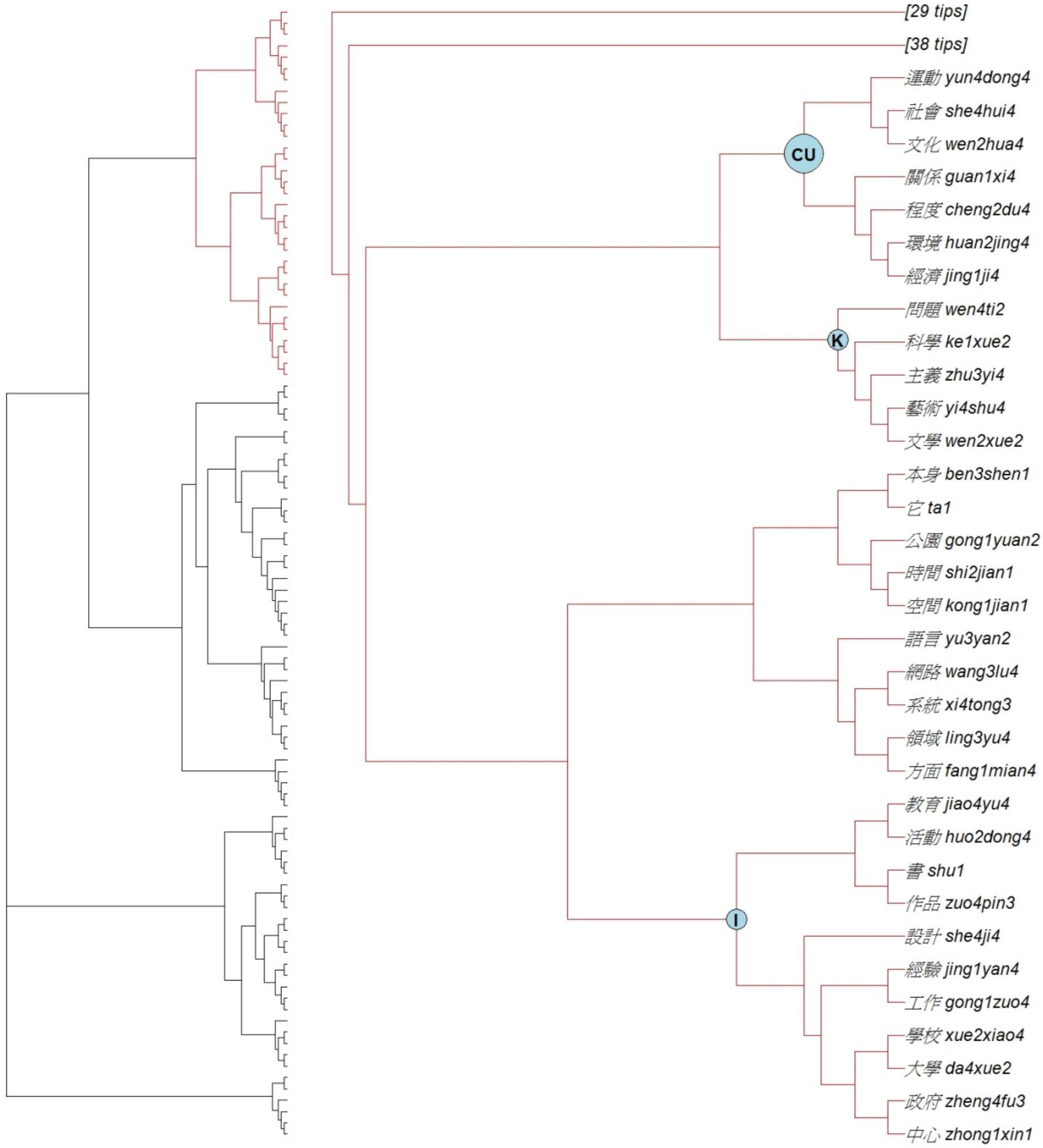


Figure 5. Subtree three of the dendrogram

Figure 6 shows the distribution of the significant covarying pairs in each semantic frame. Among all the significant covarying collexemes, about 35 percent of the NP1 falls into the HUMAN semantic frame. A further Chi-square test suggests that the distribution of these covarying pairs in different semantic frames is significant ($\chi^2_{(8)} = 862.25$, $p < 0.01$). Furthermore, among the semantic frames we identify, HUMAN presents itself as the most concrete category. This may suggest that the HUMAN frame serves as a basis for the metaphorical extension of the possessive relations encoded by MPC and that other semantic frames may be argued to derive from this basis through cognitive mechanisms.

In the semantic classes generated, however, there are still quite a range of covarying pairs that are difficult to label with appropriate semantic categories (*i.e.*, OTHER in Figure 6). Nearly one-fifth of the NP1s do not yield coherent clustering patterns at the early stage of the dendrogram. While these clusters generated in the dendrogram may not be suggestive in reaching a coherent semantic category, they are revealing in the respect that they show how one entity is conceptualized similarly to another under the context of a possessive relation. On the top of the subtree in Figure 3, it is observed that *sheng1huo2*, *sheng1ming4*, and *ren2sheng1* are often portrayed as the "end point" (*zui4hou4*) in discussing their possessed properties. Similarly, the bottom of the subtree in Figure 3 shows that the properties of the abstract entities such as world, history, times, value, and meaning are often cast in the past background as those abstract NPs (*shi4jie4*, *li4shi3*, *shi2dai4*, *jia4zhi2*, *yi4yi4*) are clustered together with *guo4qu4*. Furthermore, in the middle of the subtree in Figure 4, it is suggested that time and space is conceptualized as one coherent domain as *shi2jian1* and *kong1jian1* are clustered together at the early stage. Of similar nature is the grouping of *yu3yan2* with *wang3lu4* and *xi4tong3*, suggesting that native speakers often conceptualize the Internet and language in a similar fashion. Instead of being a blow to the credibility of our clustering method, these cases may serve as *prima facie* evidence for the degree of grammaticization in MPC toward becoming a pure "associative" syntactic formative. This paradox should not come as a counter-expectation at all to discourse-functional grammarians as the more frequent a construction gets used the more its semantics gets bleached (Hopper & Traugott, 1993; Traugott & Dasher, 2002). Yet, compared with the other 76% of the semantically coherent clusters, this small portion of the heterogamous patterns may not necessarily stop us from claiming that MPC indeed has semantic coherence in its usage.

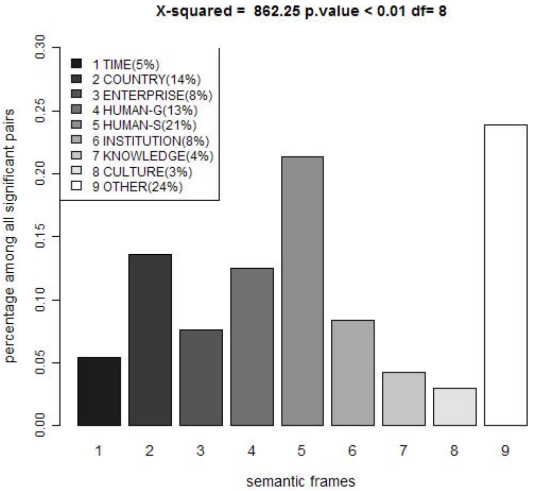


Figure 6. distribution of the significant covaryign collexemes in different semantic frames

Although the clusters are automatically yielded by the algorithm, what each cluster represents still relies on the analyst’s manual labeling, thus drawing criticism that such endeavors are still introspective and subjective. Nevertheless, it should be noted that the distribution in Figure 6 differs greatly from the raw frequency distribution used in traditional corpus linguistic studies. First of all, collostrength, rather than raw frequency, is used to reduce the possibility of making a by-chance observation. Second, even though the label for each semantic category may be analyst-dependent, the members of each cluster are objectively generated by the quantified features and a sophisticated algorithm. When adopting the same algorithm on the same dataset, different quantitative corpus linguists will obtain the same clustering results, although their labels for those semantic frames may differ. This advantage provides the possibility for research on the same construction to compare their conclusions and theoretical implications.

On the one hand, we still need a more objective way to decide what kind of semantic relations are maintained in each semantic frame. In the current stage, synsets in WordNet provide a promising possibility for an automatic identification of semantic relations (c.f., Moldovan & Badulescu, 2005). The present study only provides a coarse-grained categorization for the semantic domains of the possessor NPs. With a semantically disambiguated and syntactically parsed corpus such as WordNet, we could conduct the covarying collexeme analysis on a “synset,” rather than “word,” basis. Furthermore, clustering possessor NP1 according to possessed NP2 (or the other way around) will not provide us a clear picture of the semantic relations encoded by MPC. To automatically identify such semantic relations between NP1 and NP2, we need to cluster the whole MPC according to its covarying lexemes/constructions, such as the coocurring predicates.

On the other hand, the labeling of the semantic frame for the clusters generated may be expected to proceed automatically in the near future by making reference to the hypernyms in Chinese WordNet as well. For instance, for all the NP1s that are clustered together, we can generate a list of their hyponyms for each sense of the NP1 in WordNet and look for potential higher-order semantic domains among all these NP1s. A sophisticated extension to the synonyms of these NP1s may also facilitate the search for a common superordinate domain. In other words, the analyst's subjectivity may be reduced to the minimum once Chinese Wordnet is available. Also, it should be noted that cluster analysis here is not intended to completely substitute for manual classification (or in any sense bearing absolute superiority over the latter). Instead, the goal here is to show that, in order to introduce findings and observations from discourse-functional linguistics into the modeling of natural language processing, an automatic constructional sense induction may be needed for efficient implementation.

4.2 A Closer Look at each Semantic Frames

Before we start to look at some examples from each cluster generated, we would like to emphasize that the labels of semantic relations in the following discussion are mainly for exposition[5]. Furthermore, we leave for future consideration whether it is feasible to reach a consensus among discourse-functional grammarians regarding a unanimous set of semantic relations (See 4.3 below for more discussion). Rather, these brief sketches of the covarying collexemes in each cluster are to support our claim that the clusters generated by the neighbor-joining algorithm are indeed semantically coherent.

First of all, two types of HUMAN frames - specific and generic - can be clearly identified in Figure 4. One consists of personal pronouns while the other includes noun phrases mostly referring to the generic idea of "people" or "human beings". The former semantic frame, dubbed as HUMAN-specific, demonstrates prototypical "ownership" (*e.g.*, *ta1 DE xiao3shuo1*), "component-whole" (*e.g.*, *ta1 DE shou3*) as well as "interpersonal relation" (*e.g.*, *ta1 DE zhang4fu5*) relation between NP1 and NP2 and the covarying collexemes of higher collostrength are included in (1)[6]. In the latter, the HUMAN-generic frame, NP2 often refers to the key components of a human life or human beings in general, thus maintaining a component-whole relation with the NP1. Typical examples are included in (2).

(1) HUMAN - specific (H-s)

她 ta1 'she'	丈夫 zhang4fu5 'husband'
她 ta1 'she'	女兒 nu3er2 'daughter'
他 ta1 'he'	小說 xiao3shuo1 'novel'
他 ta1 'he'	太太 tai4tai5 'wife'
我 wo3 'I'	心 xin1 'heart'
我 wo3 'I'	心情 xin1qing2 'mood'
他 ta1 'he'	手 shou3 'hand'
我 wo3 'I'	日記 ri4ji4 'diary'

[5] Our semantic relations are based on a more complete list of semantic relations proposed by Moldovan *et al.* (2004).

[6] The covarying collexemes listed as examples here are all of significant collostrength ($p < 0.01$).

(2) Humans - generic (H-g)

人 ren2 'man'	一生 yi1sheng1 'all one's life'
人 ren2 'man'	天性 tian1xing4 'nature'
自己 zi4ji3 'self'	生命 sheng1ming4 'life'
人民 ren2min2 'people'	生活 sheng1huo2 'life'
個人 ge4ren2 'individual'	自由 zi4you2 'freedom'
我們 wo3men5 'we'	社會 she4hui4 'society'
我們 wo3men5 'we'	孩子 hai2zi5 'child'
我們 wo3men5 'we'	祖先 zu3xian1 'ancestor'
自己 zi4ji3 'self'	家 jia1 'home'
人類 ren2lei4 'humanity'	理性 li3xing4 'sense'
人 ren2 'man'	尊嚴 zun1yan2 'dignity'

In Figure 3, three semantic frames are identified: TIME, COUNTRY, and ENTERPRISE. Typical significant covarying collexemes in the TIME frame are included in (3). The purpose of this Time frame appears to contextually "position" the NP2 within a specific temporal space denoted by NP1. Therefore, it can be observed that the prominent semantic relation is attribute-holder between NP1 and NP2.

(3) TIME (T)

當時 dang1shi2 'then'	心情 xin1qing2 'mood'
今天 jin1tian1 'today'	主題 zhu3ti2 'theme'
當時 dang1shi2 'then'	台灣 Tai2wan1 'Taiwan'
現在 xian4zai4 'modern'	年輕人 nian2qing1ren2 'young people'
目前 mu4qian2 'at the present time'	狀況 zhuang4kuang4 'condition'

For the COUNTRY frame, significant covarying collexemes are listed in (4). The components of a country are clearly shown in the covarying collexemes of this category as component-whole relation appears to be a dominant semantic relation in this semantic frame. Quite a range of fundamental components of a country manifest clearly, from concrete entities like *min2zhong4* or *ren2min2* to more abstract assets such as *zheng4zhi4*, *jing1ji4*, *wen2hua4*, and *fa3lu4*. As far as the purpose of the present study is concerned, this COUNTRY frame may be argued to exhibit a metaphorical conceptualization, where a basic possessive relation -

component-whole - is extended to a higher abstract level of political entities. Also, the prominence of this semantic category may reflect the nature of the material collected in Academia Sinica Corpus as local news accounts for the majority of the data sources.

(4) COUNTRY (CO)

地區 di4qu1 'area'	人民 ren2min2 '(the) people'
國家 guo2jia1 'country'	人民 ren2min2 '(the) people'
台灣 Tai2wan1 'Taiwan'	主權 zhu3quan2 'sovereignty'
台灣 Tai2wan1 'Taiwan'	民主 min2zhu3 'democracy'
當地 dang1di4 'local'	民俗 ming2shu2 'customs'
地區 di4qu1 'area'	民眾 min2zhong4 'people'
當地 dang1di4 'local'	居民 ju1min2 'resident'
國家 guo2jia1 'country'	法律 fa3lu:4 'law'
台灣 Tai2wan1 'Taiwan'	政治 zheng4zhi4 'politics'
美國 Mei3guo2 'America'	軍事 jun1shi4 'military affairs'
台灣 Tai2wan1 'Taiwan'	原住民 yuan2zhu4min2 'indigenous peoples'
大陸 da4lu4 'mainland'	煤 mei2 'coal'
大陸 da4lu4 'mainland'	經濟 jing1ji4 'economy'
日本 Ri4ben3 'Japan'	經濟 jing1ji4 'economy'

Let us now consider the ENTERPRISE frame, as illustrated in Figure 3. There is quite a bit noise in this group, where NP1 and NP2 may hold an ownership relation (*gong1si1 DE lao3ban3*), or producer-product (*gong1si1 DE chan3pin3*) and some other typical behaviors or expectations of a social institution (*shi4chang3 DE gong1xu1* and *shi4chang3 DE jing4zheng1*). Nonetheless, this may suffice as to argue that an ENTERPRISE frame is emergent from our daily uses of MPC as all these possessor NPs (NP1) bear great resemblance in reference with their possessed entities (NP2). Furthermore, the amalgamation of *wei4lai2* with this ENTERPRISE cluster may suggest that in the discourse context these enterprises are often cast as futuristic entities in that possibilities and potentials are more emphasized.

(5) ENTERPRISE (E)

未來 wei4lai2 'future'	方向 fang1xiang4 'direction'
未來 wei4lai2 'future'	走向 zou3xiang4 'trend'
未來 wei4lai2 'future'	主人翁 zhu3ren2weng1 'master (of one's own destiny, etc.)'
市場 shi4chang3 'market'	主流 zhu3liu2 '(n) main stream of a fluid'
公司 gong1si1 '(business) company'	老闆 lao3ban3 'boss'
產品 chan3pin3 'goods'	行銷 xing2xiau1 'marketing'
市場 shi4chang3 'market'	佔有率 zhan4yau3lu4 'percentage of coverage'
市場 shi4chang3 'market'	供需 gong1xu1 'supply and demand'
公司 gong1si1 '(business) company'	股東 gu3dong1 'stockholder'
產品 chan3pin3 'goods'	品質 pin3zhi4 'quality'
公司 gong1si1 '(business) company'	董事長 dong3shi4zhang3 'chairman of the board'
市場 shi4chang3 'market'	競爭 jing4zheng1 'to compete'

In the subtree of Figure 5, three more semantic frames are identified: CULTURE, KNOWLEDGE, and INSTITUTION. Typical examples of the first frame are illustrated in (6). The NP1 in the CULTURE frame often refers to the products out of our socialization, such as *she4hui4*, *wen2hua4*, and *yun4dong4*. Typical cases of a component-whole relation in this frame may include *she4hui4 DE cheng2yuan2*, *yun4dong4 DE chuan4shi3ren2*, or *wen2hua4 DE ren2qun2*. Of particular interest here is that most possessive relations maintained between these covarying collexemes are also deemed metaphorical in the sense that the possessor and the possessed refer to abstract social entities, rather than concrete animate subjects.

(6) CULTURE (CU)

社會 she4hui4 'society'	成員 cheng2yuan2 'member'
運動 yun4dong4 'movement'	創始人 chuang4shi3ren2 ‘founder'
文化 wen2hua4 'culture'	人群 ren2qun2 'a crowd'
社會 she4hui4 'society'	良心 liang2xin1 'conscience'
社會 she4hui4 'society'	現象 xian4xiang4 'appearance'
文化 wen2hua4 'culture'	精髓 jing1sui3 'marrow'
文化 wen2hua4 'culture'	影響 ying3xiang3 'influence'
文化 wen2hua4 'culture'	差異 cha1yi4 'difference'
文化 wen2hua4 'culture'	產物 chan3wu4 'product'

Another semantic frame identified in Figure 5 - KNOWLEDGE - illustrates possessive relations in a variety of knowledge-based domains, such as *ke1xue2*, *zhu3yi4*, *yi4shu4,* and *wen2xue2*. Of particular interest here is the inclusion of *wen4ti2* into this cluster. In other words, the former disciplines such as *ke1xue2*, *zhu3yi4*, and *yi4shu4* may be argued to behave similarly to *wen4ti2* under the context of making references to their possessed entities (*i.e.*, NP2). This amalgamated pattern may suggest that the other disciplines in this KNOWLEDGE frame are often viewed as a question to which we quest for a possible solution or answer.

(7) KNOWLEDGE (K)

科學 ke1xue2 'science'	方法 fang1fa3 'method'
問題 wen4ti2 'problem'	方法 fang1fa3 'method'
主義 zhu3yi4 'creed'	色彩 se4cai3 'tint'
藝術 yi4shu4 'art'	形式 xing2shi4 'form'
問題 wen4ti2 'problem'	時候 shi2hou5 'time'
藝術 yi4shu4 'art'	創作 chuang4zuo4 'to create'
問題 wen4ti2 'problem'	答案 da2an4 'answer'
問題 wen4ti2 'problem'	辦法 ban4fa3 'means'
問題 wen4ti2 'problem'	關鍵 guan1jian4 'crucial'
問題 wen4ti2 'problem'	癥結 zheng1jie2 'bottleneck'
文學 wen2xue2 'literature'	性格 xing4ge2 'nature'
科學 ke1xue2 'science'	知識 zhi1shi5 'intellectual'

The final semantic frame - INSTITUTION - refers to goal-oriented social formations, ranging from concrete entities like *xue2xiao4*, *da4xue2*, and *shu1*, to more abstract ones like *zhong1xin1*, *huo2dong4*, and *jiao4yu4*. In terms of basic possessive relations, NP2 in this frame often consists of the components of NP1 such as *zhong1xin1 DE ren2yuan2*, *xue2xiao4 DE lao3shi1*, *shu1 DE zuo2zhe3*, *zheng4fu3 DE fa3ling4*, *da4xue2 DE xiao4zhang3*, and *xue2xiao4 DE she4bei4*. Nonetheless, a look at the NP2 shared by the NP1 in this frame suggests the goal-oriented nature of this category, as in *zheng4fu3 DE zhu3zhang1*, *jiao4yu4 DE mu4di4, hau2dong4 DE mu4di4,* and *shu1 DE zhu3zhi3*. More examples are listed in (8).

(8) INSTITUTION (I)

中心 zhong1xin1 'center'	人員 ren2yuan2 'staff'
學校 xue2xiao4 'school'	老師 lao3shi1 'teacher'
書 shu1 'book'	作者 zuo2zhe3 'author'
政府 zheng4fu3 'government'	法令 fa3ling4 'decree'
大學 da4xue2 'university'	校長 xiao4zhang3 'president'
學校 xue2xiao4 'school'	設備 she4bei4 'equipment'
活動 huo2dong4 'activity'	內容 nei4rong2 'content'
書 shu1 'book'	內容 nei4rong2 'content'
教育 jiao4yu4 'to educate'	內容 nei4rong2 'content'
政府 zheng4fu3 'government'	主張 zhu3zhang1 'to advocate'
活動 huo2dong4 'activity'	目的 mu4di4 'purpose'
教育 jiao4yu4 'to educate'	目的 mu4di4 'purpose'
政府 zheng4fu3 'government'	決策 jue2ce4 'decision'
書 shu1 'book'	主旨 zhu3zhi3 '(n) gist'

4.3 Raw Frequency and Collostrength

As the rank-ordering of the raw frequency has been greatly utilized in the literature of traditional corpus linguistic studies, we would now like to express some issues with the validity of this approach. In order to examine the relationship between the raw frequency (*i.e.*, the counts of the covarying collexemes in our collected sample) and the collostrength (*i.e.*, the association strength of the covarying collexemes with each other in the construction), we compare the ordering of these two measures for the most frequent N covarying collexemes. The procedure is as follows. First, the most frequent N covarying collexemes are selected and their corresponding raw frequency and collostrength are submitted to Friedman's rank test to see if the rank-ordering of the raw frequency and the collostrength differs significantly among these top frequent N cases. The results are shown in Table 5.

Table 5. The p-values from Friedman's rank test and Kendall's τ coefficient for the ordering of raw frequency and collostrength among the top frequent N covarying collexemes

For top frequent N covarying collexemes	Friedman test *p*-value	Kendall's τ
3	0.083265	1
4	0.0455	0.666667
5	0.025347	0.4

6	0.014306	0.466667
7	0.008151	0.52381
8	0.004678	0.357143
9	0.0027	0.055556
10	0.001565	0.022222
11	0.000911	0.163636
12	0.003892	0.090909
13	0.002282	0.102564
14	0.001341	0.230769
15	0.000789	0.314286
16	0.000465	0.383333
17	0.000275	0.441176
18	0.000162	0.503268
19	9.60E-05	0.54386
20	5.70E-05	0.463158

Table 5 illustrates the correlation between the raw frequency and the collostrength for the most frequent N covarying collexemes. The second column lists the *p*-value from Friedman test and the third column gives the Kendall's τ coefficient as the degree of correspondence between the two rankings of raw frequency and collostrength. As can be seen, while raw frequency may have explanatory power in the topmost frequent cases, the rank ordering itself may be legitimately applied only to the most frequent cases (N < 7). Starting from the most frequent 7 covarying collexemes, the rank-ordering of the raw frequency differs significantly from that of the collostrength ($\chi^2_{(1)} = 7$, p-value < 0.01). Furthermore, Kendall's τ coefficient shows the association strength of the rankings between raw frequency and the collostrength weakens with the inclusion of more covarying collexeme types. In other words, a study based on the most frequent 6 covarying collexemes may yield the same conclusions as one based on the covarying collexemes of the top 6 collostrength. Nonetheless, a study based on more than 6 covarying collexemes is likely to yield somewhat different patterns from one based on a more statistically sophisticated measure, *i.e.*, collostrength. Whether the index for rank-ordering is statistically sophisticated may be trivial for the most frequent few cases. Yet, as far as the majority of the covarying collexemes are concerned, the statistical sophistication of the rank-ordering index is non-trivial and crucial in drawing conclusions. Nevertheless, what most traditional corpus-based studies do is to base their theorizing on the rank ordering of the raw frequency in all cases, which in our view may seriously undermine the validity of such corpus-based endeavor. Therefore, we suggest that a certain level of sophistication is needed in the use of the raw frequency in traditional corpus-based studies.[7]

Let us now take a closer look at the differences between the ordering of raw frequency and that of collostrength. Table 6 shows the top 20 covarying collexemes sorted by their raw frequency in a descending order. If an analyst bases a study on the ordering of the raw frequency, they may easily reach the conclusion that the possessors in MPC overwhelmingly fall into the human category. Nevertheless, the high frequency of the covarying pairs in table 6 may derive from the fact that those NP1 are indeed words of high frequency in the overall corpora. If the frequency of the NP1 is high, the pairs containing NP1 are expected to be higher. In other words, the significance of the high constructional frequency may be diminished by the frequency of its parts. Most importantly, it remains unclear whether the observed frequency is significantly higher than the expected.

Table 6. A list of covarying collexemes ranked by their respective raw frequency

NP1	NP2	N	Collostrength
我 wo3 'I'	心 xin1 'heart'	152	101.9261
我們 wo3men5 'we'	社會 she4hui4 'society'	110	81.6629
自己 zi4ji3 'self'	生活 sheng1huo2 'life'	102	27.86079
他 ta1 'he'	作品 zuo4pin3 'works (of art)'	100	40.82442
我 wo3 'I'	孩子 hai2zi5 'child'	90	41.54748
我 wo3 'I'	手 shou3 'hand'	78	40.03977
他 ta1 'he'	話 hua4 'dialect'	77	28.52847
自己 zi4ji3 'self'	身體 shen1ti3 '(human) body'	77	46.91627
人 ren2 'man'	生命 sheng1ming4 'life'	72	49.52095
她 ta1 'she'	手 shou3 'hand'	67	46.39019
自己 zi4ji3 'self'	生命 sheng1ming4 'life'	66	28.51615
月 yue4 'moon'	時間 shi2jian1 'time'	66	94.69845
我 wo3 'I'	朋友 peng2you5 'friend'	65	28.84368
他 ta1 'he'	人 ren2 'man'	63	1.48E-05
他 ta1 'he'	朋友 peng2you5 'friend'	61	21.59611
他 ta1 'he'	手 shou3 'hand'	58	19.41221

[7] For a thorough review of statistical measures of association, please refer to Chapter 20 in Jurafsky and Martin (2008 [2000]).

自己 zi4ji3 'self'	孩子 hai2zi5 'child'	55	17.00112
自己 zi4ji3 'self'	能力 neng2li4 'ability'	55	16.34834
我們 wo3men5 'we'	生活 sheng1huo2 'living'	55	15.38543
我 wo3 'I'	意思 yi4si5 'idea'	51	28.90842
我 wo3 'I'	話 hua4 'dialect'	51	14.85332
類型 lei4xing2 'type'	人 ren2 'man'	51	51.11585
他 ta1 'he'	心 xin1 'heart'	49	9.045286
他 ta1 'he'	生活 sheng1huo2 'life'	49	1.698984
方面 fang1mian4 'respect'	問題 wen4ti2 'problem'	48	29.34759
我們 wo3men5 'we'	孩子 hai2zi5 'child'	48	23.34222
我 wo3 'I'	眼睛 yan3jing1 'eye'	47	24.80487
我 wo3 'I'	人 ren2 'man'	46	1.54E-06
他 ta1 'he'	父親 fu4qin1 'father'	46	25.52748
你 ni3 'you'	忠告 zhong1gao4 'advice'	45	79.65025

Even though we have adopted collostrength of the covarying collexemes as a reference or approximation to their association to the construction, we still do not know what kind of semantic relations MPC encodes most often. A traditional corpus linguist may proceed to label the semantic relations between NP1 and NP2 manually. In order to demonstrate a traditional corpus linguistic approach, we take the top 20 covarying collexeme pairs as an illustration. Table 7 shows the top 20 covarying collexeme pairs that are significantly attracted to each other in MPC. The list is ranked according to their collostrength in a descending order. In the rightmost column, we manually label these significant covarying collexemes with possible semantic profiles, *i.e.*, a semantic relation between NP1 and NP2. Our labels for the semantic relations in Table 7 are purely descriptive, as stated in Section4.2; no theoretical significance is attached to the precise labels used to characterize the semantic relations.

Table 7. A list of covarying collexemes ranked by their respective collostrength

NP1	NP2	N	Collostrength	Semantic relation
不久 bu4jiu3 'not long (after)'	將來 jiang1lai2 'future'	35	107.7124	Idiom
我 wo3 'I'	心 xin1 'heart'	152	101.9261	Component-Whole
月 yue4 'moon'	時間 shi2jian1 'time'	66	94.69845	Attribute-Holder
我們 wo3men5 'we'	社會 she4hui4 'society'	110	81.6629	ownership
你 ni3 'you'	忠告 zhong1gao4 'advice'	45	79.65025	Participant-Event
政府 zheng4fu3 'government'	政策 zheng4ce4 'policy'	40	59.58043	Participant-Event
魔王 mo2wang2 'fiend'	左手 zuo3shou3 'left-hand'	14	51.66222	Component-Whole
類型 lei4xing2 'type'	人 ren2 'man'	51	51.11585	Attribute-Holder
人 ren2 'man'	生命 sheng1ming4 'life'	72	49.52095	ownership
最後 zui4hou4 'final'	獵人 lie4ren2 'hunter'	19	49.42944	idiom
媒體 mei2ti3 'media'	報導 bao4dao3 'coverage'	24	49.00789	Participant-Event
自己 zi4ji3 'self'	身體 shen1ti3 '(human) body'	77	46.91627	Component-Whole
她 ta1 'she'	手 shou3 'hand'	67	46.39019	Component-Whole
問題 wen4ti2 'problem'	癥結 zheng1jie2 'bottleneck'	20	42.58891	Component-Whole
龍 long2 'dragon'	傳人 chuan2zen2 'heir'	12	41.89946	idiom
我 wo3 'I'	孩子 hai2zi5 'child'	90	41.54748	Interpersonal relations
因素 yin1su4 'element'	影響 ying3xiang3 'influence'	22	41.51503	Participant-Event
他 ta1 'he'	作品 zuo4pin3 'works'	100	40.82442	ownership
我 wo3 'I'	手 shou3 'hand'	78	40.03977	Component-whole
生命 sheng1ming4 'life	意義 yi4yi4 'meaning'	37	40.03224	Attribute-

(force)'				Holder
學者 xue2zhe3 'scholar'	社區 she4qu1 'community'	16	38.85216	ownership
異樣 yi4yang4 'discrimination'	眼光 yan3guang1 'judgment'	14	38.56742	Attribute-Holder
國王 guo2wang2 'king'	新衣 xin1yi1 'new clothes'	11	38.10723	idiom
動詞 dong4ci2 'verb'	論元 lun4yuan2 'argument'	11	38.03616	Component-Whole
人 ren2 'man'	一生 yi1sheng1 'all one's life'	35	37.90407	Participant-Event
挫折 cuo4zhe2 'setback'	時候 shi2hou5 'time'	22	36.72331	Time-Event
瘤子 liu2zi5 'lump'	老公公 lao3gong1gong1 'old man'	9	36.0871	Attribute-Holder
他 ta1 'he'	妻子 qi1zi5 'wife'	44	35.55265	Interpersonal Relations
方面 fang1mian4 'respect'	知識 zhi1shi5 'knowledge'	30	34.97364	Attribute-Holder
用戶 yong4hu4 'user'	需求 xu1qiu2 'requirement'	19	34.68546	Participant-Event

In Table 7, several covarying collexemes of low frequency do jump out as prominent instances of MPC, such as *mo2wang2 DE zuo3shou3*, *long2 DE chuan2zen2*, *guo2wang2 DE xin1yi1*, and *dong4ci2 DE lun4yuan2*. These significant pairs are not only indicative of the constructional semantic profiles but also suggestive of the topics covered in the corpora. Crucially, these phrases would not have emerged on the analyst's list if one had adopted only raw frequency as their measure of association.

Interpretable as it may seem, even the results based on the ordering of the collostrength still raise several methodological issues. Although we have flavored a traditional corpus linguistic approach with a quantitative nature using collostrength, such a traditional approach still needs to face the fact that a predetermined list of semantic relations is needed in order to label all the covarying pairs. It comes as no surprise that our labeling for the semantic relations in Table 7 may draw adverse criticism from researchers of a different paradigm. Linguists differ greatly in the number of possible semantic relations encoded by possessive constructions and different linguists may adopt different terms. For instance, the semantic relations can be summarized into 10 labels as in Stefanowitsch (2003) or can be further elaborated into 35 as in Moldovan *et al.* (2004). Furthermore, while a small sample of the significant covarying collexemes may be indicative of the basic semantic profiles of MPC, there is still a potential drawback. We choose the top 20 covarying collexemes only for demonstration of a traditional corpus linguistic approach. A traditional corpus linguistic study could have chosen the top 200, 2000, or even 20000. In other words, to the size of the sample from the ordering list may have great impact on the validity of the results. As long as a traditional corpus linguist likes to investigate all the semantic relations between NP1 and NP2 in MPC, they are bound to face these potential challenges. Most importantly, it would be difficult for them to bypass the issue of how to classify all the MPC tokens in an objective way. While a wedge of cheese like the top 20 (or more) covarying collexemes may be suggestive for the semantic coherence of MPC, a step further can be made to include more data so as to generate the semantic coherence of MPC in a more objective fashion. This is exactly the niche we are trying to occupy.

5. Concluding Remarks

Based on our empirical investigation, the overall results suggest that Mandarin Possessive Construction does exhibit a considerable degree of semantic coherence that holds between covarying collexemes, and the relative consistency among different sets of covarying collexemes. In addition, we further ensure the objectivity in identifying semantic classes of the possessor NPs by submitting a sample of covarying collexemes into phylogenic hierarchical clustering. The generated dendrogram appears to support the claim that semantic coherence does hold between covarying collexemes of the construction in question and NP1 exhibits several clear semantic classes where possessive relations are often contextualized, namely, HUMAN, COUNTRY, ENTERPRISE, INSTITUTION, KNOWLEDGE, and CULTURE. Nevertheless, some of the clusters have failed to manifest a coherent category of their own. While the prominent semantic frames identified may explain why most linguists still recognize this construction as a possessive construction in Mandarin, these heterogeneous clusters may account for the fact that some would describe it as a pure contextually-driven formative for any possible association. Therefore, noise in our results may serve as preliminary evidence for its degree of grammaticization toward becoming a pure "associative" syntactic formative.

The purpose of the present study should be clear. While construction grammar has emerged as one of the dominant theoretical frameworks in the usage-based research paradigm, its insights may be further supported by more quantitative empirical data. It is argued that covarying collexeme analyses may serve as a compelling approach in identifying constructional sub-patterns, thus lending more credibility to the empirical results. Also, various statistical tools may not only facilitate the difficult task of categorization for the analysts but reduce the subjectivity of the judgment to the minimum as well.

Furthermore, with more and more quantitative methods being incorporated into linguistic studies, these findings are more likely to be taken seriously by other interdisciplinary scholars.

Differences in methodology only widen the gap for the possible interdisciplinary interaction and comparison. Crucially, while other disciplines like biology, psychology, and cognitive science have long been viewing classification as a quantitative problem and have been using computer programs to identify a best parsimonious tree from an unorganized dataset, it would be less advantageous for traditional linguists to opt for an intuition-based approach where classifications are acceptable as long as scholars of the same research paradigm agree that they are acceptable. Even though traditional corpus linguistics has made a step further in contributing a great deal to the linguistic theorizing in general, such an approach does not typically produce data which are interpretable and usable by neighboring disciplines, especially in natural language processing. While other disciplines provide results based on rigorous quantitative design, they would hardly buy the story of linguists who generate conclusions via purely descriptive statistics. Therefore, a more rigorous quantitative method may serve as an objective platform where more interdisciplinary dialogue on human cognition can be made. While discourse-functional and cognitive linguists are sifting the wheat from the chaff in the massive harvest of corpus data, it is hoped that such rigorous emphasis on methodology may lend more objectivity and credibility to their revealing insights.

References

Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. *Systematic Biology,* 54(4), 513-526.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baron, I., Herslund, M., & Sorensen, F. (2001). *Dimensions of possession*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Biq, Y.-O. (2001). The grammaticalization of *Jiushi* and *Jiushishou* in Mandarin Chinese. *Concentric: Studies in English Literature and Linguistics,* 27(2), 53-74.

Biq, Y.-O. (2004a). Construction, reanalysis, and stance: 'V *yi ge* N' and variations in Mandarin Chinese. *Journal of Pragmatics,* 36, 1655-1672.

Biq, Y.-O. (2004b). From collocation to idiomatic expression: The grammaticalization of *hao* phrases/constructions in Mandarin Chinese. *Journal of Chinese Language and Computing,* 14(2), 73-95.

Biq, Y.-O. (2004c). People, things, and stuff: general nouns in spoken Mandarin. *Concentric: Studies in Linguistics,* 30(1), 41-64.

Bybee, J. (1998). The emergent lexicon. *Chicago Linguistic Society,* 34, 421-435.

Bybee, J. (2005). Mechanisms of change in grammaticization: The role of frequency. In B. D. Joseph & R. D. Janda (Eds.), *The Handbook of Historical Linguistics* (pp. 602-623). Malden, MA: Blackwell Publication.

Bybee, J., & Hopper, P. J. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degree of constituency: the reduction of *don't* in American English. *Linguistics, 37*, 575-596.

Chang, M.-H. (2002). Discourse functions of *Anne* in Taiwanese Southern Min. *Concentric: Studies in English Literature and Linguistics,* 28(2), 85-115.

Chui, K. (2000). Morphologization of the degree adverb *HEN*. *Language and Linguistics,* 1(1), 45-59.

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Croft, W. (2008). Evolutionary Linguistics. *Annual Review of Anthropology,* 37(1), 219-234.

Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Curran, J. R. (2004). *From distributional to semantic similarity.* Unpublished dissertation, University of Edinburgh, Edinburgh, UK.

Dancygier, B., & Sweetser, E. E. (2000). Constructions with *if*, *since* and *because*: Causality, epistemic stance, and clause order. In E. Couper-Kuhlen & B. Kortmann (Eds.), *Cause, condition, concession, contrast: Cognitive and discourse perspectives* (pp. 111-142). Berlin: Mouton de Gruyter.

Divjak, D. S., & Gries, S. T. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory,* 2(1), 23-60.

Dong, C.-R. (2003). A cognitive account of possessive construction. *Foreign Languages and Their Teaching,* 169, 60-63.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., & Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science,* 309(5743), 2072-2075.

Fillmore, C. J., & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of *risk* and its neighbors. In A. Lehrer & E. F. Kittay (Eds.), *Frames, Fields, and Contrasts* (pp. 75-102). Hillsdale, NJ: Lawrence.

Fillmore, C. J., Kay, P., & O'Connor, M. K. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language,* 64, 501-538.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument structure*. Chicago: Chicago University Press.

Goldberg, A. E. (1998). The emergence of the semantics of argument structure constructions. In B. MacWhinney (Ed.), *The emergence of language* (pp. 197-212). Hillsdale, NJ.: Lawrence Erlbaum Associates.

Goldberg, A. E., Casenhiser, D., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics,* 15, 286-316.

Grady, J. (1997). *Foundations of meaning: Primary metaphors and primary scenes.* Unpublished dissertation, University of California Berkeley, Berkeley, CA.

Gries, S. T., & Stefanowitsch, A. (2004). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, Culture and Mind* (pp. 225-236). Stanford, CA: CSLI Publications.

Gries, S. T., & Stefanowitsch, A. (to appear). Cluster analysis and the identification of collexeme classes. In J. Newman & S. Rice (Eds.), *Experimental and empirical methods in the study of conceptual structure, discourse, and language* (pp. 73-90). Stanford, CA: CSLI Publications (Available at: http://www.linguistics.ucsb.edu/faculty/stgries/research/ClusteringCollexemes.pdf).

Heine, B. (2001). Ways of explaining possession. In I. Baron, M. Herslund & F. Sorensen (Eds.), *Dimensions of possession* (pp. 311-328). Amsterdam and Philadelphia, PA.: John Benjamins.

Hilpert, M. (2007). *Germanic Future Constructions: A Usage-based Approach Grammaticalization.* Unpublished dissertation, Rice University, Houston, TX.

Hopper, P. J. (1987). Emergent grammar. *Berkeley Linguistics Society,* 13, 139-157.

Hopper, P. J., & Traugott, E. C. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

Huang, S. (1998). Emergent lexical semantics. In S. Huang (Ed.), *Selected papers from the second international symposium on languages in Taiwan* (pp. 129-150). Taipei: Crane.

Huang, S. (2003). Doubts about complementation: A functionalist analysis. *Language and Linguistics,* 4(2), 429-455.

Hunston, S., & Francis, G. (1999). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia, PA.: John Benjamins.

Jurafsky, D., & Martin, J. H. (2008 [2000]). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd edn ed.). Upper Saddle River, NJ: Prentice Hall.

Kaufman, L., & Rousseeuw, P. J. (2005 [1990]). *Finding groups in data: An introduction to cluster analysis* (2nd edn ed.). Hoboken, NJ: Wiley.

Kay, P., & Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language,* 75(1), 1-34.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (2nd edn ed., pp. 202-251). Cambridge: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Langacker, R. W. (1991). *Foundations of cognitive grammar: Descriptive application* (Vol. 2). Stanford, CA: Stanford University Press.

Langacker, R. W. (2003). Constructions in cognitive grammar. *English Linguistics,* 20, 41-83.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley and Los Angeles: University of California Press.

Liu, M.-C. (2002). *Mandarin Verbal Semantics: A Corpus-based Approach*. Taipei: Crane Publishing Co.

Lyons, C. (1986). The syntax of English genitive constructions. *Journal of Linguistics,* 22(1), 123-143.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McMahon, A., & McMahon, R. (2003). Finding Families: Quantitative Methods in Language Classification. *Transactions of the Philological Society,* 101(1), 7-55.

Michaelis, L. A. (2003). Word meaning, sentence meaning, and syntactic meaning. In H. Cuykens, R. Dirven & J. R. Taylor (Eds.), *Cognitive approaches to lexical semantics* (pp. 163-209). Berlin and New York: Mouton de Gruyter.

Michaelis, L. A., & Lambrecht, K. (1996). Toward a construction-based model of language function: The case of nominal extraposition. *Language,* 72, 215-247.

Moldovan, D., & Badulescu, A. (2005, October). *A semantic scattering model for the automatic interpretation of genitives.* Paper presented at the Human language technology conference and conference on empirical methods in natural language processing, Vancouver.

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., & Girju, R. (2004, 6 May). *Models for the semantic classification of noun phrases.* Paper presented at the HLT-NAACL Worksop on Computational Lexical Semantics, Boston, MA.

Nikiforidou, K. (1991). The Meanings of the Genitive: A Case Study in Semantic Structure and Semantic Change. *Cognitive Linguistics,* 2(2), 149-205.

Ono, T., & Thompson, S. A. (1996). Interaction and Syntax in the Structure of Conversational Discourse: Collaboration, Overlap, and Syntactic Dissociation. In E. H. Hovy & D. R. Scott (Eds.), *Computational and Conversational Discourse: Burning Issues, an Interdisciplinary Account* (pp. 67-96). Heidelberg: Springer-Verlag.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics,* 20(2), 289-290.

Pedersen, T. (1996). Fishing for exactness. *Proceedings of the SCSUG 96 in Austin, TX*, 188-200.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4(4), 406-425.

Scheibman, J. (2002). *Point of View and Grammar: Structural patterns of subjectivity in American English conversation*. Amsterdam: John Benjamins Publishing Company.

Stefanowitsch, A. (2003). Constructional semantics as a limit to grammatical alternation: The two genitives of English. In G. Rohdenburg & B. Mohndorf (Eds.), *Determinants of Grammatical Variation in English*. Berlin and New York: Mouton de Gruyter.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics,* 8(2), 209-243.

Stefanowitsch, A., & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory,* 1(1), 1-43.

Su, L. I.-w. (1998). Conversation coherence: the use of *ranhou* in Chinese spoken discourse *Collected Papers of the Second Interactional Symposium on Languages in Taiwan* (pp. 167-182). Taipei: Crane.

Su, L. I.-w. (2002). Why a construction - That is the question! *Concentric: Studies in English Literature and Linguistics,* 28(2), 27-42.

Su, L. I.-w. (2004). Subjectification and the use of the complementizer *SHUO*. *Concentric: Studies in Linguistics,* 30(1), 19-40.

Tao, H. (2003a). Toward an emergent view of lexical semantics. *Language and Linguistics,* 4(4), 837-856.

Tao, H. (2003b). A usage-based approach to argument structure: 'remember' and 'forget' in spoken English. *International Journal of Corpus Linguistics,* 8(1), 75-95.

Tao, H., & Thompson, S. A. (1994). The discourse and grammar interface: Preferred clause structure in Mandarin conversation. *Journal of the Chinese Language Teachers Association,* 29(3), 1-34.

Taylor, J. R. (1996). *Possessives in English: An Exploration in Cognitive Grammar*. Oxford: Oxford University Press.

Thompson, S. A. (2002). "Object complements" and conversation: towards a realistic account. *Studies in Language,* 26(1), 125-164.

Thompson, S. A., & Couper-Kuhlen, E. (2005). The clause as a locus of grammar and interaction. *Discourse Studies,* 7(4-5), 481-506.

Thompson, S. A., & Hopper, P. J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In J. Bybee & P. J. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 27-60). Amsterdam and Philadelphia: John Benjamins.

Traugott, E. C., & Dasher, R. B. (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Tyler, A., & Evans, V. (2003). *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge: Cambridge University Press.

Wang, Y.-F., Katz, A., & Chen, C.-H. (2003). Thinking as saying: *shuo* ('say') in Taiwan Mandarin conversation and BBS talk. *Language Sciences,* 25(5), 457-488.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory,* 4(2), 253-290.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member：	NT$ 1,000 (US$50.-)
group membership：	NT$20,000 (US$1,000.-)

life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#：________________

Name：__________________________________ Date of Birth：_______________

Country of Residence：____________________ Province/State：____________________

Passport No.：___________________________ Sex: _______________________

Education(highest degree obtained)：____

Work Experience：____

____

Present Occupation：____

Address：____

____

Email Add：____

Tel. No：______________________________ Fax No：____________________________

Membership Category：□ Regular Member □ Life Member

Date：_____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member	：	US$ 50.- （NT$ 1,000）
Life Member	：	US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號		（由本會填寫）	
姓　　名		性別		出生日期	年　月　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
		申請人:		（簽章）	
		中　華　民　國		年　　月	日

審查結果:

1. 年費：

   終身會員：　10,000.-

   個人會員：　1,000.-

   學生會員：　500.-（限國內學生）

   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)

   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638

   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw

   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

PAYMENT FORM

Name :____(Please print) Date: ___________________

Please debit my credit card as follows: US$________________________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:_________________________

Card No.: __________- __________ - __________ - __________ Exp. Date:________________

3-digit code: ______________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE : ____________________________________

Tel.: ___________________________________ E-mail: ___________________________________

Add: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (CLCLP)

Quantity Wanted: ______

US$ _____ ❑ Publications:___________________________________

US$ _____ ❑ Text Corpora: __________________________________

US$ _____ ❑ Speech Corpora:________________________________

US$ _____ ❑ Others: _______________________________________

US$ __________ ❑Life Member Fee ❑ New Member ❑Renew

US$ __________ = Total

Fax : 886-2-2788-1638 or Mail this form to :

ACLCLP
% Institute of Information Science, Academia Sinica
R502, 128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會

信用卡付款單

姓名: ________________________________(請以正楷書寫)　日期:：______________

卡別：□ VISA CARD □ MASTER CARD □ JCB CARD　發卡銀行：________________

卡號:_________-_________-_________- ________　有效日期：_________________

卡片後三碼：_____________________（卡片背面簽名欄上數字後三碼）

持卡人簽名：　__________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：_______________________　E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________□ 中文計算語言學期刊(IJCLCLP)

NT$__________□ 中研院詞庫小組技術報告

NT$__________□ 中文（新聞）語料庫

NT$__________□ 平衡語料庫

NT$__________□ 中文詞庫八萬目

NT$__________□ 中文句結構樹資料庫

NT$__________□ 平衡語料庫詞集及詞頻統計

NT$__________□ 中英雙語詞網

NT$__________□ 中英雙語知識庫

NT$__________□ 語音資料庫______________________________

NT$__________□ 會員年費　□續會　□新會員　□終身會員

NT$__________□ 其他: ___________________________________

NT$_______________=　合計

填妥後請傳真至 02-27881638 或郵寄至:
115台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

			AIR	AIR		
		Surface	(US&EURP)	(ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： □ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期） 年份：______ （過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：______________________________ 收據抬頭：______________________________

地　　址：____

電　　話：______________________________ E-mail: ______________________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.

1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.

2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.

3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.

4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript

6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.

7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

> Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

> Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).

Please visit the following websites for details.

(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.14 no.3 September 2009

Contents

Papers

章無疵也章之明靡句無玷也句之清英字不妄也文賦曰選義按部考辭就班就所傳達者觀之禮記曰發志為言[illegible]言為名傳曰言以足志文[illegible]足言易曰書不盡言言不盡意詩序曰在心為志發言為詩情動於中而形於言蓋情志發而[illegible]成語言工而文字傳也

ISSN: 1027-376X
The Association for Computational Linguistics and Chinese Language Processing