WORLD CHINESES BASED ON COMPARABLE CORPUS:
The Case of Grammatical Variations of jinxing

Chu-Ren Huang, Jingxia Lin, Huarui Zhang

Abstract: The term World Chineses (全球華語), though not as common as World Englishes, is becoming more and more widely used with the increasing popularity of Chinese as a second language and with the Chinese diaspora spreading and growing. The lexical variations among World Chineses are easily observed and often studied. Yet, to better understand the dynamicity of World Chineses and facilitate efficient communication among speakers of World Chineses, we need to study grammatical variations in the different varieties of Chinese. In this study, we propose a comparable-corpus-based approach for grammatical comparison. With the case study of the construction of the light verb 進行 jinxing 'proceed' in Taiwan and Mainland Mandarin, we demonstrate that our approach makes it possible a systematic and comprehensive study of grammatical variations of World Chineses.

Keywords: Grammatical Variations, Comparable Corpora, World Chineses, jinxing construction

1. Introduction

Mandarin Chinese is the most widely-spoken language in the world. Despite the same linguistic heritage that enables mutual intelligence in general, Mandarin spoken in different regions has evolved in different ways as a result of the political, economic, cultural, and social development of each region. Comprehensive documentation and explanation of the variations will facilitate better communication among the Chinese-speaking communities and help resolve any misunderstanding due to such variations.

In recent years, a variety of studies have been carried out for lexical variations such as region-specific neologism, meaning variations of the same word, and the use of different words to express the same meaning. For instance, several dictionaries have been compiled to list the lexical differences among Chinese-speaking communities, especially those between Mainland and Taiwan: Qiu (1990), Wei & Sheng (2000), Shi et al. (2003), Li (2010). A number of case studies have been carried out on the regional variations of lexicon, too. For instance, Tang (1995) and Zhao (2008) provide examples showing how a concept is expressed differently in Mainland, Taiwan, Hong Kong, and Singapore Mandarin. Wang & Li (1996), D. Xu (1995), X. Xu (1995), Zeng (1995), among many others, discuss the lexical differences between Mainland and Taiwan Mandarin. Chen (2000), Zhou (2008), Shi & Zhu (2005), and Shi et al. (2010) analyze the different words used in Hong Kong and Mainland Mandarin. Zhou (1999), Wang (1999), and Xing (2005) compare the lexicons in Singapore and Mainland Mandarin.

While many of the studies introduced above are based on researchers' introspection, a number of lexical comparisons have also been carried out based on corpus and even statistical technology. For instance, Huang et al. (2007) and Simon et al. (2007) adopt bootstrapping co-occurrence statistics from tagged and segmented Chinese corpus to automatically identify diverging transliterations of foreign named entities into Mainland and Taiwan Mandarin. Hong & Huang (2008), among others, develop methodologies to detect the lexicons and their collocations that are used differently in Mainland and Taiwan Mandarin. In addition, corpus-based lexical comparisons have been conducted on topics such as thesaurus (Kwong & Tsou 2006, 2008), Chinese news headlines (Chin 2007), celebrity coverage in the Chinese press (Tsou et al. 2005), and judgment terms (Kwong & Tsou 2004).

However, unlike lexical comparison, seldom research has been undertaken on grammatical variations, especially those based on comparable corpora. The previous studies can be found in Lu (1995, 1996, 2002) on the grammatical differences between Mainland and Singapore Mandarin, and Shi et al. (2005), Shi et al. (2006), Shi & Wang (2006), among others, on Mainland and Hong Kong Mandarin. Nevertheless, none of these studies are based on large-scale corpora.

We argue that the lack of comprehensive and systematic studies to a large extent is due to the nature of grammatical variation. Studies of grammatical variations target structurally larger units such as phrases, constructions, and sentences, and investigate whether diversity exists in distributional and selectional constraints on how these unites are constructed. Compared with lexical variation, grammatical variation is more complex and thus, technically more difficult to be automatically detected. Furthermore, grammatical variation is subtle in that there are no grammatically “correct” variants versus “incorrect” variants, but grammatical variants that are more conventionalized and favored to use versus those that are grammatical but less used.

We propose the comparable-corpus-based approach for grammatical comparison. In the rest of this paper, we introduce the features and merits of the approach and the methodology of constructing comparable corpora in Section 2. In Section 3, with the case study of 進行 ‘proceed’ in Taiwan and Mainland Mandarin, we demonstrate that the comparable-corpus-based approach is able to identify grammatical variations that may not be easily found by traditional comparative studies. Conclusions are drawn in Section 4.

2. Comparable Corpus

According to Eagles (1996), a comparable corpus is composed of “similar texts in more than one language or variety”. Eagles points out that comparable corpus such as the International Corpus of English (Greenbaum 1991) selects texts from the languages or varieties based on the same criteria such as size, genre, and time. It differs from parallel corpora in that the latter consist of corpora translated into different languages. The most important advantage of using large-scale comparable corpus is that it enables the comparison of different languages or varieties in “similar circumstance of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus” (Eagles 1996) or by direct quotation of data from another language or variety. In this sense, comparative studies based on a comparable corpus can best objectively discover the differences in the usage of the constructions under investigation.

We propose that comparable corpora can be built through dynamic re-construction. For different linguistic purposes, dynamic re-construction can be carried out in different ways. Temporal, spatial, topical, and grammatical re-constructions are four common types. The first two types are relatively easier to understand, that is, the re-reconstructions are extraction and collection of corpus data based on time (e.g., data from a particular year or month) or space (e.g., data from certain locations). Topical re-construction extracts and collects data from a base corpus according to certain topics, e.g., flood, Chinese president, cancer. Grammatical re-construction searches grammatically annotated data (e.g., part of speech, collocation) and extracts phrases, constructions, or sentences for further comparison. The overall similarity of the sub-corpora in a comparable corpus and individual dissimilarity of grammatical unit or pattern can be statistically measured and thus serve as the quantitative criteria for constructing comparable corpus (see Binary Distributional Consistency in Zhang et al. 2004). Compared with the traditional way of sub-corpus building that is limited to time, space, genre, or other less complicated parameters, dynamic re-construction is able to achieve more sophisticated purposes in a more reliable and finer-grained way.

During recent years, several large-scale corpora of contemporary Mandarin Chinese have been constructed, which thus provides a basis for dynamic re-construction. For instance, the Chinese Gigaword corpus has collected over 1.1 billion Chinese words, consisting of data from Central News Agency (Taiwan, about 700 million characters), Xinhua News Agency (Mainland China, about 400 million characters), and Zaobao Newspaper (Singapore, about 30 million characters) (Huang 2006, Hong & Huang 2006). The LIVAC corpus (Linguistic Variations in Chinese Speech Communities) is a synchronous corpus with 400 million Chinese characters, containing representative media texts from Hong Kong, Taiwan, Beijing, Shanghai, Macau and Singapore over 16 years (Tsou et al. 2011). The PKU-CCL corpus constructed by the Center for Chinese Linguistics at Peking University collects texts of different genres in Modern Chinese, with a size of over 300 million characters[1]. The availability of such corpora makes it possible for the construction of comparable corpora for different linguistic purposes.

3. Case Study: The construction of the light verb jinxing

Light verb in this paper refers to semantically impoverished verbs that may contribute information about event shape (e.g., beginning or ending of an event), but specify little about event structure (cf. Zhu 1999, Diao 2004, among others). The predicative content of a light verb construction comes from the event-denoting element that is taken as complement by the light verb. For instance, in the construction 進行討論 proceed-discuss 'discuss', the event of discussion is denoted by the complement 討論, whereas the light verb 進行 specifies an inceptive meaning of the event. While many light verbs take derived event nominals and event NPs as their complements, variations can exist in the semantic and syntactic types of the complements in different Mandarin-speaking communities. For instance, 從事 'undertake' in Taiwan Mandarin can take complements denoting negative events such as 性交易 'sex trade' and 勾當 'shady business', whereas such complements are seldom found with 從事 in Mainland Mandarin.

In our case study, we explore the grammatical differences of the light verb 進行 in Taiwan and Mainland Mandarin, as well as how the variations are related to the argument structure of the verb. Particularly, we will focus on the grammatical status of the event-denoting elements that can be taken as complements, i.e. whether the complement can be an event noun (半決賽 'semifinal'), a deverbal event noun (討論 discuss-discuss 'discussion'), or even a verbal phrase (驗票 inspect-ticket 'inspect votes'). Through the case study, we illustrate the key role that comparable corpora play in the research of grammatical variations. The rest of this section first introduces previous observation of 進行, followed by our comparable-corpus-based study.

3.1 Previous study of jinxing construction

Previous studies can be found for both Mainland and Taiwan 進行, especially the constraints on the complements taken by 進行. (1) is a summary of the studies on the features of the complements to Mainland 進行.

(1) a. Must denote durative events (Lü 1982, Lu 2000)
進行尋找 'proceed to look for' vs *進行找到 'proceed to find (successfully)'

b. Must denote formal events (Lü 1982)
進行談話 'proceed to have a conversation' vs *進行說話 'proceed to talk'

c. Cannot be monosyllabic (Lü 1982, Lu 2000)
進行教導 'proceed to teach' vs *進行教 'proceed to teach'

d. Cannot take another object (Lü 1982)
進行幫助 'proceed to help' vs *進行幫助他 'proceed to help him'

e. Must denote spontaneous/controllable events (Lu 2000)
進行摧毀 'proceed to destroy' vs *進行坍塌 'proceed to collapse'

f. Can only be coordinated VV complements, e.g., 進行教學 'proceed to teach', but not VC, VO, SV complements, e.g., *進行教課 'proceed to teach (a lesson)' (Luo & Feng 2010)

g. Can be nouns (戰爭 'war'), noun phrases (第二項議程 'the second item on the agenda'), verbs (計算 'compute'), VO phrases (定論 draw-conclusion) (Diao 2004)[2]

進行 in Taiwan Mandarin has similar uses to its counterpart in Mainland Mandarin. For instance, according to Huang et al. (1995) and Liu & Chang (2005), 進行 is found in formal register and takes complements that denote durative and atelic events. Liu & Chang (2005) find that in Sinica Corpus, 進行 tends to select verbs of conversation that are often used in formal style, e.g., 討論 'discuss' and 溝通 'communicate', but not 談論/吵架/聊天 'talk/quarrel/chat' which are less formal.

Furthermore, previous comparative studies, e.g., Wei & Sheng (2000) and Shi et al. (2003), indicate no difference between Mainland and Taiwan 進行.

3.2 A comparable-corpus-based study of jinxing construction

In this section, we first present the comparable corpora for the extraction of 進行 constructions. Then, we compare and analyze the complements that occur in these constructions. The results show that while Mainland and Taiwan 進行 share a large number of complements, the latter differs from the former in that it also tends to take more types of NPs and even VPs as its complements.

3.2.1 Data

We selected data from the Gigaword corpus (Section 2) to construct comparable corpora based on the parameter of time (the corpora are also comparable in terms of space because the data is either from Mainland or Taiwan Mandarin). As illustrated in Table 1, a Taiwan and a Mainland corpus from the year of 2000 are collected respectively to constitute a comparable corpus; in addition, a second comparable corpus is constructed with Taiwan and Mainland Mandarin data from Year 2004. In other words, four sub-corpora were used for the study; they are Taiwan 2000, Mainland 2000, Taiwan 2004, and Mainland 2004.

Table 1 Comparable corpus by temporal re-construction

	Total number of words	
	Taiwan	Mainland
Gigaword 2000	46 million	20 million
Gigaword 2004	27 million	28 million

Table 1 shows that the sizes of the sub-corpora are not balanced: Taiwan 2000 (46 million words) is about twice the size of Mainland 2000 (20 million words), whereas Taiwan 2004 (27 million words) is about the same size of Mainland 2004 (28 million words). Furthermore, the numbers of tokens of 進行 found in each sub-corpus, as given in Table 2, are also not proportionate to the sizes of the sub-corpora: Mainland 2004 has the largest number of 進行 although its size is not the largest of the four sub-corpora.

Table 2 Tokens of 進行 in each sub-corpus of the comparable corpora

	Tokens of 進行	
	Taiwan	Mainland
Gigaword 2000	43,000	34,000
Gigaword 2004	23,000	45,000

However, in terms of relative frequency of occurrence, 進行 in both Mainland 2000 and Mainland 2004 is about 1.8 times more frequently used than in Taiwan 2000 and Taiwan 2004, as illustrated in Figure 1. This indicates that there is a difference in the frequency of occurrence of 進行 in the two varieties of Mandarin.

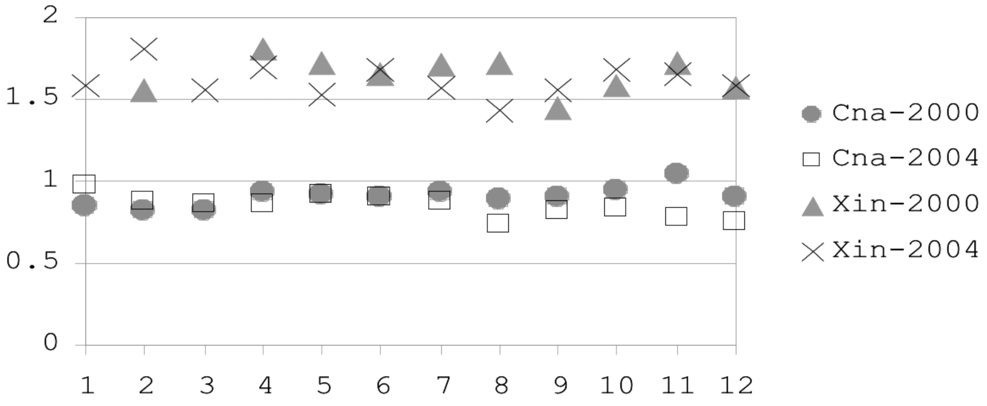

Figure 1 Relative frequencies of Taiwan and Mainland 進行
(Y-axis: relative frequency; X-axis: month)

3.2.2 Extraction of *jinxing* construction

We adopted the context-free template approach, i.e. regular expression, to extract 進行 and its complements. In order to make the linguistic analysis feasible, only the complements with a minimal frequency of five are selected for further analysis. Table 3 lists the number of types of complements to both Taiwan and Mainland 進行.

Table 3 Number of types of complements to Taiwan and Mainland 進行

	Total types of complement	Types unique to Taiwan	Types shared by Taiwan and Mainland	Types unique to Mainland
Gigaword 2000	415	180	104	131
Gigaword 2004	405	121	126	158

Table 3 indicates that although 進行 in the two varieties share complements, each also has complements that the other does not have. In addition, there are more types unique to Taiwan 進行 in Year 2000, but more types unique to Mainland 進行 in Year 2004.

A summary of the types of complements is presented in Table 4.

Table 4 Types of complements to 進行

Types of complement		Taiwan Only		Share		Mainland Only	
		2000	2004	2000	2004	2000	2004
Derived event nominal	V+V (討論 discuss-discuss‘discussion’)	114	78	84	106	90	111
	Adv+V (直播 direct-broadcast‘live broadcase’)	21	18	8	10	13	22
	V+O (修憲 amend-constitution ‘constitutional amendment’)	18	13	7	3	12	14
Simple noun (決賽 ‘final contest’)		20	8	5	7	11	6
VP (驗票 inspect-ticket ‘inspect votes’)		5	3	0	0	0	0
Other		2	1	0	0	5	5
Sum		180	121	104	126	131	158

The complements to 進行 can be summarized into three major types: derived event nominal, simple noun, and VP. In our study, derived event nominal refers to any linguistic unit that describes a situation (Grimshaw 1990) and still has verbal use in Modern Chinese. For instance, 討論 discuss-discuss ‘discussion/discuss’ is marked as an event nominal when it occurs after 進行 (2a), but it also can be used as a verb, as in (2b).

(2) a. 每三週與規劃小組的老師進行一次討論 (Gigaword Taiwan)
‘[The student team] has a discussion with the teachers of the planning team every three weeks.’
b. 雙方討論了緊迫的國際問題 (Gigaword Taiwan)
‘The two sides discussed urgent international issues.’

The notion of “simple noun” (e.g., 決賽 ‘final’) refers to nouns that also describe a situation (cf. nouns denoting an entity such as 蘋果 ‘apple’ and 桌子 ‘table’). But unlike derived event nominal, a simple noun never has a verbal use. In contrast, “VP” (e.g., 驗票 inspect-ticket ‘inspect votes’) describes a situation and in general only has verbal uses.

Table 4 indicates that a V-O combination can be a derived event nominal (e.g., 修憲 amend-constitution) or a VP (e.g., 驗票 inspect-ticket). According to our definitions, if a V-O is found with both noun and verbal uses, it is marked as an event nominal when it is taken by 進行 as a complement; in contrast, if a V-O is in most cases used as a verbal phrase, then it is marked as a VP. In other words, a V-O combination is an event nominal if it can be used as a noun. Therefore, the V-O combinations that can be modified by a relative clause or a sequence of “(determinative) numeral/ quantifier + classifier” are event nominals. For instance, 修憲 is an event nominal as it can appear in constructions such as 一次修憲 ‘a constitutional amendment’ and 歷時三年的修憲 ‘a constitutional amendment that cost three years’, whereas 抽水 behaves like a VP because it is unnatural to occur in a noun position, as in ? 一次抽水 ‘a water pump’, ? 歷時三個小時的抽水 ‘a water pump that cost three hours’.

Table 4 shows that there are both similarities and differences between the complements to Taiwan and Mainland 進行. Derived event nominal is the most common type of complement to both Mainland and Taiwan 進行. In addition, V-V event nominal is found to be the largest type. Table 5 lists more examples of the V-V nominals. It indicates that a large number of the nominals are shared by 進行

in Mainland and Taiwan, e.g., 分析 divide-analyze ‘analysis/ analyze’, 治療 cure-cure ‘treatment/cure’, 試驗 test-test ‘experiment/to experiment’, and 操作 operate-operate ‘operation/operate’.

Table 5 V-V derived event nominals taken as complements to Taiwan and Mainland 進行

	Taiwan only	Shared	Mainland only
Gigaword 2000	比對；抗爭；拆除；辯論；研議；瞭解；干擾；攻防；查察；抗議；演練；救援；etc.	協調；分析；準備；治療；試驗；查處；操作；競爭；修改；比較；偵查；往來；etc.	保護；建設；打擊；救治；培養；威脅；捕撈；教育；獎勵；管理；鬥爭；解釋；etc.
Gigaword 2004	比對；抗爭；拆除；辯論；研議；瞭解；評論；監督；說明；追蹤；掃蕩；追蹤；etc.	協調；分析；準備；治療；試驗；查處；操作；抗議；演練；救援；解釋；鬥爭；etc.	保護；建設；打擊；救治；培養；威脅；學習；安排；聯繫；恢復；指責；修理；etc.

Besides the similarities, Taiwan 進行 differs from its Mainland counterpart in terms of the types of simple nouns and VPs it can take as complements. We elaborate on the differences in the rest of this section.

Table 4 suggests that both Taiwan and Mainland 進行 can take some simple nouns as their complements. However, more types of simple nouns are found with Taiwan 進行. A list of the complements is provided in Table 6.

Table 6 Simple nouns taken as complements to Taiwan and Mainland 進行

	Taiwan only	Shared	Mainland only
Gigaword 2000	民調；X之旅 (感恩之旅，學術之旅)；君子之爭；票選；環評；情蒐；安檢；X程 (議程，工程，賽程)；	X賽 (決賽); 手術	和談
Gigaword 2004	民調；疫調；X會議；賽程 (最後一天賽程)；X事宜 (合作事宜)；X程序 (準備程序)；言詞辯論庭；此事	X賽 (決賽)；手術；戰爭；和談；體檢；貿易	None

Table 6 shows that while event nouns such as 決賽 'final', 手術 'operation', and 戰爭 'war' are found with both 進行, the Taiwan variety also collocates with nouns such as 感恩之旅 'thanksgiving trip', 君子之爭 'gentle dispute', 票選 'ballot election' and 環評 'Environmental Impact Assessment'. Furthermore, nouns such as 言詞辯論庭 'oral argument tribunal' and 此事 'this event/activity' that do not directly denote an event are also found with Taiwan 進行, as illustrated in (3).

(3) a. 我們 必須 平靜 的 進行 此 事。 (Gigaword Taiwan)
'We must deal with the issue/carry out the activity calmly.'
b. 司法院 大法官 會議 今天 繼續 進行 言詞 辯論 庭。(Gigaword Taiwan)
'Today, the Grand Justices Council continued oral argument in the tribunal.'
(literary, 'continued proceeding oral argument tribunal')

The other significant difference between Taiwan and Mandarin 進行 is their ability to take VP complements. As illustrated in Table 7, while Mainland 進行 is not found with any VP complements, Taiwan 進行 can collocate with VPs, with a few examples given in (4).

Table 7 VPs taken as complements to Taiwan and Mainland 進行

	Taiwan only	Shared	Mainland only
Gigaword 2000	抹黑；驗票；測謊；喊話；抽水	None	None
Gigaword 2004	驗票；開票；配票	None	None

(4) a. 利用 已 真相大白 的 事情 進行 抹黑 ， 企圖 影響 選情 (Gigaword Taiwan)
'[He] took advantage of the truth to discredit [Qiu], in an attempt to influence the election.'
b. 明天下午投票後，進行開票。(Gigaword Taiwan)
'Ballot counting will be started after the vote tomorrow afternoon.'

c. 宣傳車 大聲 播放 音樂 和 進行 喊話 (Gigaword Taiwan)
'Propaganda vehicles played music and conduct propaganda loudly.'

In order to verify that Mainland 進行 indeed does not favor VP complements, we also searched for the VP complements to Taiwan 進行 (as in Table 7) in two larger corpora of Mainland Mandarin: the whole Gigaword Mandarin corpus (data in Year 1991-2004, approximately 400 million characters) and the CCL-PKU Corpus (approximately 300 million characters). As illustrated in Table 8, still very few VPs are found to collocate with Mainland 進行 in the two larger corpus, which thus is consistent with the results in Table 7.

Table 8 VPs taken as complements to Mainland 進行 in Gigaword and CCL-PKU corpus

VP complement found unique to Taiwan 進行 in Gigaword 2000 and 2004	Frequency of occurrence as complement to Mainland 進行 in CCL-PKU corpus	Frequency of occurrence as complement to Mainland 進行 in Gigaword Mainland Chinese
X 開票	0	1
X 驗票	5	4
X 喊話	3	0
X 抹黑	2	2
X 測謊	2	11[4]
X 配票	0	0

(note: "X" refers to the possible modifiers before the VP)

To summarize, our comparable-corpus-based case study finds that Mainland and Taiwan 進行 share the similarities that both favor VV derived event nominals as their complements, which is consistent with the observation of previous studies such as Lü (1982), Lu (2000), and Diao (2004). However, Mainland Mandarin speakers may find an "overuse" of 進行 by Taiwan speakers because the latter can select more types of NP complements and even VPs as its complement (cf. Wei & Sheng

2000, Shi et al. 2003, Diao 2004, Luo & Feng 2010, among others). The finding of the different selectional restriction of 進行 in the two varieties contributes to the documentation of grammatical variations. In addition, the collocational features of 進行 found in this study can also be served as the annotation reference for dynamic grammatical re-construction of comparable corpora in the future.

Furthermore, the diversity and data of this case study can also be used as the foundation for future linguistic research. For instance, studies are necessary to identify the possible factors that determine the types of complements collocating with 進行, as well as the reasons why the differences exist. In addition, diachronic study is required to investigate the development of VP as the complements to 進行. One possible hypothesis is that VP was Commonly taken as complements by 進行 as derived event nominals were in earlier stages of Mandarin, but now has disappeared in Mainland Mandarin and is gradually dying away from Taiwan Mandarin. The other possible hypothesis is that while VP was not allowed in a 進行 construction in the past, it is now making incursions into the complement to Taiwan 進行 and possibly will appear in Mainland 進行 in future.

4. Conclusions

In this paper, we propose a comparable-corpus-based approach for the studies of grammatical variations in Mandarin Chinese. Through the case study of the constructions of the light verb 進行, we show that such an approach is able to discover the variations that are not observed by previous traditional studies and thus, provides empirical basis for a systematic and comprehensive documentation and explanation of grammatical variations of World Chinese.

Note:

[1] http://ccl.pku.edu.cn:8080/ccl_corpus/

[2] Note that although Diao (2004) claims that the complement to Mainland 進行 can be VO phrases, the examples (eg, 定論 draw-conclusion ‘final conclusion’) he provides are

event nominals rather than real VO phrases, as they usually do not have verbal uses in Mainland Chinese.

[3] If we google Internet data from Mainland China, the VPs that are unique to Taiwan 進行 can also be found in 進行 constructions. However, it is observed that most of these examples are quoted or reproduced from Taiwan texts (especially news). In this sense, these Internet examples cannot be used as evidence to show that Mainland 進行 is compatible with these VPs. The different results from the searches in comparable corpus and Internet also indicate that the former is more reliable for comparative studies, whereas the latter must be used with caution.

[4] Compared with other VPs, 測謊 occurs most frequently as the complement to Mainland 進行 (11 instances), which indicates a possible collocational relationship between 進行 and 測謊.

References

陳建民 2000 《內地與香港的詞語比較》，《語文建設》第4期。

刁晏斌 2004 《現代漢語虛義動詞研究》， 遼寧師範大學出版社。

洪嘉馡、黃居仁 2008 《語料庫為本的兩岸對應詞彙發掘》，《 語言暨語言學》第9期。

盧福波 2000《對外漢語常用詞語對比例釋》，北京語言大學出版社。

陸儉明 1995《關於新加坡華語規範化問題》，《新加坡聯合早報》1995年6月16日。

——— 1996《新加坡華語語法的特點》，《新加坡南大中華語言文化學報》。

——— 2002《新加坡華語句法特點及其規範問題》，《新馬華人傳統與現代的對話》，南洋理工大學。

羅艷娟、馮莉娟 2010《淺談“進行”的賓語》，《 語文學刊》第7期。

呂叔湘 1999 [1980]《現代漢語八百詞》，商務印書館。

丘質樸 1990《大陸和台灣差別詞典》，南京大學出版社。

石定栩、劉藝、盛玉麒 2010《香港書面漢語常見自造詞研究》，《詞彙學理論與應用》第5輯，商務印書館。

石定栩、邵敬敏、朱志瑜 2006《港式中文與標準漢語的比較》，香港教育圖書公司。

石定栩、王冬梅 2006《香港書面漢語的語法特點》，《中國語文》第2期。

石定栩、朱志瑜 2005《英語對香港書面漢語詞彙的影響－香港書面漢語和標準漢語中的同形異義詞》，《外國語》第6期。

石定栩、朱志瑜、王冬梅 2005《香港書面漢語體標記的特點》，《雙語雙方言》，漢學出版社。

施光亨、李行健、李鑾 2003《兩岸現代漢語常用詞典》，北京語言大學出版社。

湯志祥　1995《中國大陸，台灣，香港，新加坡漢語詞彙方面若干差異舉例》，《徐州師範學院學報》。

王鐵昆、李行健 1996《兩岸詞彙比較研究管見》，《華文世界》第81期。

魏勵、盛玉麒 2000《大陸及港澳台常用詞對比詞典》，北京工業大學出版社。

徐丹暉　1995《兩岸詞語差異之比較》，《第一屆兩岸漢語語彙文字學術研討會論文集》。

許學仁　1995《海峽兩岸新詞語中若干詞義衍生和規範的考察》，《第一屆兩岸漢語語彙文字學術研討會論文集》。

曾榮汾　1995《兩岸語言詞彙整理之我見》，《第一屆兩岸漢語語彙文字學術研討會論文集》。

趙一凡 2008《淺談兩岸三地同實異形詞及其規范問題》，《現代語文（語言研究版）》第3期。

周歲嵬 2008《內地與香港詞彙的差異與融合》，《長春教育學院學報》第24期。

朱德熙　1999《現代書面漢語的虛化動詞和名動詞》，《朱德熙文集(3)》，商務印書館。

Chin, Andy 2007 A sociocultural comparison of Chinese news headline verbs between Hong Kong and Taiwan, *Proceedings of The First International Workshop on Intercultural Collaboration - IWIC 2007*, 444-450.

EAGLES 1996 Preliminary recommendations on corpus typology.
http://www ilc cnr it/EAGLES/corpustyp/corpustyp html (accessed September 30 2011)

Greenbaum, Sidney 1991 The development of the International Corpus of English In Karin Aijmer and Bengt Altenberg (eds), *English Corpus Linguistics: Studies in honour of Jan Svartvik*, 83-91. London: Longman

Grimshaw, Jane 1990 *Argument Structure*. Cambridge: MIT Press.

Hong, Jia-Fei and Chu-Ren Huang 2006 Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research, *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation.*

Huang, Chu-Ren 2006 Automatic Acquisition of Linguistic Knowledge: From Sinica Corpus to Gigaword Corpus, Invited paper at the *13th National Institute of Japanese Language International Symposium Language Corpora: Their Compilation and Application, Tokyo*, March 6-7.

Huang, Chu-Ren, Meili Yeh, and Li-Ping Chang 1995 A corpus-based study of nominalization and verbal semantics: Two light verbs in Mandarin Chinese, *Proceedings of the Sixth North American Conference on Chinese Linguistics*.

Huang, Chu-Ren, Petr Simon and Shu-Kai Hsieh 2007 Automatic Discovery of Named Entity Variants, *Proceedings of the Association of Computational Linguistics Annual Meeting*, 153-156.

Kwong, Oi Yee and Benjamin Tsou 2004 The usage and perception of judgement terms in the Pan-Chinese context, *Proceeding of 5th Chinese Lexical Semantics Workshop* (CLSW5), 220-227.

----- Regional variation of domain-specific lexical items: Toward a Pan-Chinese lexical resource, *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 9-16.

Kwong, Oi Yee and Benjamin K Tsou 2008 Extending a thesaurus with words from Pan-Chinese sources, *Proceedings of the 22nd International Conference on Computational Linguistics*.

Liu, Mei-Chun and Chun Edison Chang 2005 From frame to subframe: Collocational asymmetry in Mandarin verbs of conversation *Computational Linguistics and Chinese Language Processing*, 10(4): 431-444.

Simon, Petr, Chu-Ren Huang, Shu-Kai Hsieh and Jia-Fei Hong 2007 Transliterated Named Entity Recognition Based on Chinese Word Sketch. *International Journal of Computer Processing of Oriental Languages* 21(1): 19-30.

Tsou, Benjamin K, Andy C Chin and Oi Yee Kwong 2011 From synchronous corpus to monitoring corpus, LIVAC: The Chinese case, *DBKDA 2011: The Third International Conference on Advances in Databases, Knowledge, and Data Applications*, 175-180.

Tsou, Benjamin, Raymond Yuen, Oi Yee Kwong, Tom, Lai and Wong and Wei Lung Wong 2005 Polarity classification of celebrity coverage in the Chinese press, *Proceedings of 2005 International Conference on Intelligence Analysis*.

Zhang, Huarui, Chu-Ren Huang and Shiwen Yu 2004 Distributional Consistency: As a general method for defining a core lexicon, *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC2004)*, 1119-1122.

(**Chu-Ren Huang, Jingxia Lin:** The Hong Kong Polytechnic University;
Huarui Zhang: The Hong Kong Polytechnic University/ Peking University**)**

基於可比語料庫的全球華語研究芻議：以“進行”結構的語法變異為例

黄居仁　林靜夏　張化瑞

提要：隨著漢語作為第二語言的不斷普及以及海外華人的日益擴散和增長，“世界華語”這個名詞也越來越被廣泛使用。目前已有大量關於世界華語詞彙變異的研究。然而為了深入地理解世界華語的動態性及幫助漢語使用者克服差異，我們也需研究世界華語的語法變異。本文指出語法變異的研究應該基於可比語料庫。通過對輕動詞“進行”的個案研究，我們證明基於可比語料庫的方法可以使世界華語全面系統的研究成為可能。

關鍵詞：世界華語；語法變異；可比語料庫；“進行”結構