Regional Variation of Domain-Specific Lexical Items: Toward a Pan-Chinese Lexical Resource

Oi Yee Kwong and Benjamin K. Tsou
Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{rlolivia,rlbtsou}@cityu.edu.hk

Abstract

This paper reports on an initial and necessary step toward the construction of a Pan-Chinese lexical resource. We investigated the regional variation of lexical items in two specific domains, finance and sports; and explored how much of such variation is covered in existing Chinese synonym dictionaries, in particular the Tongyici Cilin. The domain-specific lexical items were obtained from subsections of a synchronous Chinese corpus, LIVAC. Results showed that 20-40% of the words from various subcorpora are unique to the individual communities, and as much as 70% of such unique items are not yet covered in the Tongyici Cilin. The results suggested great potential for building a Pan-Chinese lexical resource for Chinese language processing. Our next step would be to explore automatic means for extracting related lexical items from the corpus, and to incorporate them into existing semantic classifications.

1 Introduction

Many cities have underground railway systems. Somehow one takes the *tube* in London but the *subway* in New York. In a more recent edition of the Roget's Thesaurus (Kirkpatrick, 1987), *subway*, *tube*, *underground railway* and *metro* are found in the same semicolon-separated group under head *624 Way*. Similarly if one looks up WordNet (http://wordnet.princeton.edu; Miller et al., 1990), the synset to which *subway* belongs also contains the words *metro*, *tube*, *underground*, and *subway system*; and it is further indicated that "in Paris the subway system is called the 'metro' and in London it is called the 'tube' or the 'underground'". Such regional lexical variation is also found in Chinese. For instance, the subway system in Hong Kong, known as the Mass Transit Railway or MTR, is called 地鐵 in Chinese. The subway systems in Beijing and Shanghai, as well as the one in Singapore, are also known as 地鐵, but that in Taipei is known as 捷運. Their counterpart in Japan is written as 地下鉄 in Kanji. Such regional variation, as part of lexical knowledge, is important and useful for many natural language applications, including natural language understanding, information retrieval, and machine translation. Unfortunately, existing Chinese lexical resources often lack such comprehensiveness.

To fill this gap, Tsou and Kwong (2006) proposed a comprehensive Pan-Chinese lexical resource, based on a large and unique synchronous Chinese corpus as an authentic basis for lexical acquisition and analysis across various Chinese speech communities. For a significant world language like Chinese, a useful lexical resource should have maximum *versatility* and *portability*. It is not sufficient to target at one particular community speaking the language and thus cover only language usage observed from that particular community. Instead, such a lexical resource should document the core and universal substances of the language on the one hand, and also the more subtle variations found in different communities on the other. As is evident from the above example on the variation of *subway*, regional variation should be captured for the lexical resource to be useful in a wide range of applications.

In this study, we investigate and compare the regional variation of lexical items from two spe-

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 9–16,
Sydney, July 2006. ©2006 Association for Computational Linguistics

cific domains, finance and sports, as an initial and necessary step toward the more important undertaking of building a Pan-Chinese lexical resource. In addition, we make use of an existing Chinese synonym dictionary, the *Tongyici Cilin* (Mei et al., 1984) as leverage, and explore its coverage of such variation and thus the potential for enriching it. The lexical items under study were obtained from a synchronous Chinese corpus, LIVAC, which will be further introduced in Section 4. Corpus data from four Chinese speech communities were compared with respect to their commonality and uniqueness, and also against Cilin for their coverage. Results showed that 20-40% of the words extracted from the corpus are unique to the individual communities, and as much as 70% of such unique items are not yet covered in Cilin. It therefore suggests that the synchronous corpus is a rich source for mining region-specific lexical items, and there is great potential for building a Pan-Chinese lexical resource for Chinese language processing.

In Section 2, we will briefly review existing resources and related work. Then in Section 3, we will briefly outline the design and architecture of the Pan-Chinese lexical resource proposed by Tsou and Kwong (2006). In Section 4, we will further describe the Chinese synonym dictionary and the synchronous Chinese corpus used in this study. The comparison of their lexical items will be discussed in Section 5. Future directions will be presented in Section 6, followed by a conclusion.

2 Existing Resources and Related Work

The construction and development of large lexical resources is relying more and more on corpus-based approaches, not only as a result of the increased availability of large corpora, but also for the authoritativeness and authenticity allowed by the approach. The Collins COBUILD English Dictionary (Sinclair, 1987) is amongst the most well-known lexicographic fruit based on large corpora.

For natural language applications, much of the information in conventional dictionaries targeted at human readers must be made explicit. Lexical resources for computer use thus need considerable manipulation, customisation, and supplementation (e.g. Calzolari, 1982). WordNet (Miller et al., 1990), grouping words into synsets and linking them up with relational pointers, is probably the first broad coverage general computational lexical database. In view of the intensive time and effort required in resource building, some researchers have taken an alternative route by extracting information from existing machine-readable dictionaries and corpora semi-automatically (e.g. Vossen et al., 1989; Riloff and Shepherd, 1999; Lin et al, 2003).

Compared to the development of thesauri and lexical databases, and research into semantic networks for major languages such as English, similar work for the Chinese language is less mature. This gap was partly due to the lack of authoritative Chinese corpora as a basis for analysis, but has been gradually reduced with the recent availability of large Chinese corpora including the LIVAC synchronous corpus (Tsou and Lai, 2003) used in this work and further described below, the Sinica Corpus (Chen et al., 1996), the Chinese Penn Treebank (Xia et al., 2000), and the like.

An important issue which is seldom addressed in the construction of Chinese lexical databases is the problem of versatility and portability. For a language such as Chinese which is spoken in many different communities, different linguistic norms have emerged as a result of the individualistic evolution and development of the language within a particular community and culture. Such variations are seldom adequately reflected in existing lexical resources, which often only draw reference from one particular source. For instance, Tongyici Cilin (同義詞詞林) (Mei et al., 1984) is a thesaurus containing some 70,000 Chinese lexical items in the tradition of the Roget's Thesaurus for English, that is, in a hierarchy of broad conceptual categories. First published in the 1980s, it was based exclusively on Chinese as used in post-1949 Mainland China. Thus for the subway example above, the closest word group found is 火車, 列車 (train) only, let alone the subway itself and its regional variations.

With the recent availability of large corpora, especially synchronous ones, to construct an authoritative and timely lexical resource for Chinese is less distant than it was in the past. A large synchronous corpus provides authentic examples of the language as used in a variety of locations. It thus enables us to attempt a comprehensive and in-depth analysis of the core common language in constructing a lexical resource; and to incorporate useful information relating to location-sensitive linguistic variations.

3 Proposal of a Pan-Chinese Thesaurus

The Pan-Chinese lexicon proposed by Tsou and Kwong (2006) is expected to capture not only the core senses of lexical items but also senses and uses specific to individual Chinese speech communities.

The lexical database will be organised into a core database and a supplementary one. The core database will contain the core lexical information for word senses and usages which are common to most Chinese speech communities, whereas the supplementary database will contain the language uses specific to individual communities, including "marginal" and "sublanguage" uses.

A network structure will be adopted for the lexical items. The nodes could be sets of near-synonyms or single lexical items (in which case synonymy will be one type of links). The links will not only represent the paradigmatic semantic relations but also syntagmatic ones (such as selectional restrictions).

We thus begin by investigating in depth the regional variation of lexical items, especially domain-specific words, among several Chinese speech communities. In addition, we explore the potential of enriching existing resources as a start. In the following section, we will discuss the Tongyici Cilin and the synchronous Chinese corpus used in this study in greater details.

4 Materials and Method

4.1 The Tongyici Cilin

The Tongyici Cilin (同義詞詞林) (Mei et al., 1984) is a Chinese synonym dictionary, or more often known as a Chinese thesaurus in the tradition of the Roget's Thesaurus for English. The Roget's Thesaurus has about 1,000 numbered semantic heads, more generally grouped under higher level semantic classes and subclasses, and more specifically differentiated into paragraphs and semicolon-separated word groups. Similarly, some 70,000 Chinese lexical items are organized into a hierarchy of broad conceptual categories in the Tongyici Cilin. Its classification consists of 12 top-level semantic classes, 94 sub-classes, 1,248 semantic heads and 3,925 paragraphs.

4.2 The LIVAC Synchronous Corpus

LIVAC (http://www.livac.org) stands for Linguistic Variation in Chinese Speech Communities. It is a synchronous corpus developed by the Language Information Sciences Research Centre of the City University of Hong Kong since 1995 (Tsou and Lai, 2003). The corpus consists of newspaper articles collected regularly and synchronously from six Chinese speech communities, namely Hong Kong, Beijing, Taipei, Singapore, Shanghai, and Macau. Texts collected cover a variety of domains, including front page news stories, local news, international news, editorials, sports news, entertainment news, and financial news. Up to December 2005, the corpus has already accumulated about 180 million character tokens which, upon automatic word segmentation and manual verification, amount to over 900K word types.

For the present study, we make use of the subcorpora collected over the 9-year period 1995-2004 from Hong Kong (HK), Beijing (BJ), Taipei (TW), and Singapore (SG). In particular, we focus on the *financial news* and *sports news* to investigate the commonality and uniqueness of the lexical items used in these specific domains in the various communities. We also evaluate the adequacy of the Tongyici Cilin in terms of its coverage of such domain-specific terms especially from the Pan-Chinese perspective, and thus assess the room for its enrichment with the synchronous corpus. Table 1 shows the sizes of the subcorpora used for this study.

Subcorpus	**Overall** (rounded to nearest 0.01M)		**Financial News** (rounded to nearest 1K)		**Sports News** (rounded to nearest 1K)	
	Word Token	**Word Type**	**Word Token**	**Word Type**	**Word Token**	**Word Type**
HK	14.39M	0.22M	970K	38K	1041K	39K
BJ	11.70M	0.19M	232K	20K	443K	28K
TW	12.32M	0.20M	254K	22K	657K	33K
SG	13.22M	0.21M	621K	28K	998K	34K

Table 1 Sizes of individual subcorpora

4.3 Procedures

Word-frequency lists were generated from the financial and sports subcorpora from each individual community. For each resulting list, the steps below were followed to remove irrelevant items and retain only the potentially useful content words:

(a) Remove all numbers and non-Chinese words.

(b) Remove all proper names, including those annotated as personal names, geographical names, and organisation names. Proper names have been annotated in the corpora during the process of word segmentation.

(c) Remove function words.

(d) Remove lexical items with frequency 5 or below.

The numbers of remaining items in each subcorpus after the above steps are listed in Tables 2 and 3 for the two domains respectively. The lexical items retained, which are expected to contain a substantial amount of content words, are potentially useful for the current study. The lists in each domain (from the various subcorpora) were compared in terms of the items they share and those unique to individual communities. Their unique items were also compared against the Tongyici Cilin to investigate its adequacy and explore how it might be enriched with the synchronous corpus.

Subcorpus	All	After (a)	After (b)	After (c)	After(d)
HK	37,525	27,937	20,422	17,162	5,238
BJ	20,025	17,361	14,460	12,134	2,791
TW	22,142	19,428	16,316	13,496	3,088
SG	28,193	22,829	16,863	13,822	3,836

Table 2 Number of word types remaining after various data cleaning steps for the financial domain

Subcorpus	All	After (a)	After (b)	After (c)	After(d)
HK	39,190	35,720	25,289	21,502	6,316
BJ	27,971	26,049	19,799	16,598	3,878
TW	32,706	30,231	20,361	17,248	4,601
SG	34,040	31,974	19,995	16,780	5,120

Table 3 Number of word types remaining after various data cleaning steps for the sports domain

5 Results and Discussion

5.1 Lexical Items from LIVAC

The four subcorpora of the financial domain differ considerably in their sizes, and slightly less so for the sports domain. Despite this, we observed for both domains from Tables 2 and 3 that in general about 40-50% of all word types are numbers, non-Chinese words, proper names, and function words. Of the remaining items, about 20-30% have frequency greater than 5. These several thousand word types from each subcorpus are expected to be amongst the more interesting items and form the “candidate sets” for further investigation.

5.2 Commonality among Various Regions

Comparing the candidate sets from various subcorpora, which reflect the use of Chinese in various Chinese speech communities, Tables 4 and 5 show the sizes of the intersection sets among different places for the two domains respectively.

The intersection set for all four places contains slightly more than 1,000 lexical items in the financial domain. A quick skim through these common lexical items suggests that they contain, on the one hand, the many general concepts in the financial domain (e.g. 公司 company, 市場 market, 銀行 bank, 投資 invest / investment, 業務 business, 發展 develop / development, 集團 corporation, 股份 stock shares, 股東 shareholder, 資金 capital, etc.); and on the other hand, many reportage and cognitive verbs often used in news articles (e.g. 表示 express, 認爲 reckon, 出現 appear, 反映 reflect, etc.).

In the sports domain, more than 1,700 lexical items were found in all of the four subcorpora. Like its financial counterpart, we found many general concepts at the top of the list (e.g. 球員 player, 球隊 team, 賽事 match, 比賽 competition, 聯賽 league, 教練 coach, 對手 opponent, 冠軍 champion, etc.).

The numbers of overlaps in Tables 4 suggest that lexical items used in Mainland China (as evident from BJ data) seem to have the least in common with the rest. For instance, compared to the overlap amongst all four regions (i.e. 1,039), the overlap has increased most when BJ was not included in the comparison; and when we compare any two regions, the overlap between BJ and TW is smallest. Nevertheless, such uniqueness of BJ data is less apparent in the sports domain. In particular, the difference between HK/BJ and BJ/TW is even slightly less than that in the financial domain.

If we look at the individual regions, HK apparently shares most (about 50%) with SG, and vice versa (about 68%), in the financial domain. At the same time, BJ also shares more with HK than with the other two regions, and so does TW. But surprisingly, BJ has over 60% overlap with SG and about 55% with TW in the sports domain. The overlaps of TW with HK and with BJ differ by more than 20% in the finance domain, but only by about 10% in the sports domain. All these patterns might suggest lexical items in the financial domain are more versatile and have more varied focus in different communities, whereas those in the sports domain reflect the more common interests of different places.

Regions	Overlap	Proportion to individual lists (%)			
		HK	BJ	TW	SG
HK / BJ / TW / SG	1039	19.84	37.23	33.65	27.09
HK / BJ / TW	1126	21.50	40.34	36.46	
HK / BJ / SG	1327	25.33	47.55		34.59
HK / TW / SG	1581	30.18		51.20	41.21
BJ / TW / SG	1092		39.13	35.36	28.47
HK / BJ	1609	30.72	57.65		
HK / TW	1912	36.50		61.92	
HK / SG	2607	49.77			67.96
BJ / TW	1250		44.79	40.48	
BJ / SG	1505		53.92		39.23
TW / SG	1795			58.13	46.79

Table 4 Commonality amongst various regions for the financial domain

Regions	Overlap	Proportion to individual lists (%)			
		HK	BJ	TW	SG
HK / BJ / TW / SG	1668	26.41	43.01	36.25	32.58
HK / BJ / TW	1782	28.21	45.95	38.73	
HK / BJ / SG	2047	32.41	52.78		39.98
HK / TW / SG	2249	35.61		48.88	43.93
BJ / TW / SG	1864		48.07	40.51	36.41
HK / BJ	2318	36.70	59.77		
HK / TW	2693	42.64		58.53	
HK / SG	3305	52.33			64.55
BJ / TW	2124		54.77	46.16	
BJ / SG	2554		65.86		49.88
TW / SG	2709			58.88	52.91

Table 5 Commonality amongst various regions for the sports domain

5.3 Uniqueness of Various Regions

Next we compared the lists with respect to what they have unique to themselves. Table 6 shows the numbers of unique items found in each list, together with examples from the most frequent 20 unique items in each case.

Again, taking the size difference among the candidate sets into account, about 40% of the lexical items found in HK data are unique to the region, which re-echoes the versatility and wide coverage of interests of HK data. This is especially evident when compared to only about 20% of the candidate sets for SG are unique to Singapore.[1]

Looking at the unique lexical items found in individual regions, it is not difficult to see the region-specific lexicalisation of certain concepts. For instance, in terms of housing, 居屋 (housing under the Home Ownership Scheme) is a specific kind of housing in Hong Kong, 組屋 is a specific term in Singapore (as seen in SG data), whereas housing is generally expressed as 住房 in Mainland China (as seen in BJ data). Similarly, 操練 (HK) and 冬訓 (BJ) both refer to training, but may relate to different practice in the two communities. Such regional variation lends strong support to the importance of a Pan-Chinese lexical resource.

The lists of unique items also suggest the various focus and orientation in different Chinese speech communities. For example, while Hong Kong pays much attention to the real estate market and stock market, Mainland China may be focusing more on the basic needs like water, farming, poverty alleviation, etc., and Singapore is relatively more concerned with local affairs like port management. The passion for baseball, among other more popular sports like soccer, is most obvious from the unique lexical items found in TW data.

5.4 Comparison with Tongyici Cilin

As mentioned earlier, the Tongyici Cilin contains some 70,000 lexical items under 12 broad semantic classes, 94 subclasses, and 1,428 heads. It was first published in the 1980s and was based on lexical usages mostly of post-1949 Mainland China. In this section, we discuss the results obtained from comparing the unique lexical items found from individual subcorpora with Cilin, which are shown in Table 7.

On the one hand, Cilin's collection of words may be considerably dated and obviously will not include new concepts and neologisms arising in the last two decades. On the other hand, the data in LIVAC come from newspaper materials in the 1990s. So overall speaking, for each of the unique word lists, much less than 50% are covered in Cilin.

Nevertheless, there is still an apparent gap between Cilin's coverage of the unique items from various places. About 40% of the unique items found in BJ for both domains are covered; but for other places, the coverage is more often less than 30% in either or both domains. Again, this could be considered a result of Cilin's bias toward lexical usages in Mainland China.

In addition, while almost 40% of the unique items in BJ data are found in Cilin, many of these unique items covered are amongst the most frequent items. On the contrary, even though about 560 unique items in HK data are also found in Cilin, only 3 out of the 20 most frequent items are amongst them. In addition, the apparent coverage does not necessarily suggest the correct match of word senses. For instance, 居屋 is found under head *Bn1* together with other items like 住房, 住宅, etc., all of which only refer to the general concept of housing, instead of the housing specifically under the Home Ownership Scheme as known in Hong Kong. Also, coverage of words like 晨曦, 帝王 and 水手 in the sports domain does not match their actual usages which refer to team names. A more interesting example might be 火鍋, which is used in the basketball context in TW data, and in no way refers to the literal "hot pot" sense.

Results from the above comparisons thus support that (1) different Chinese speech communities have their distinct usage of Chinese lexical items, in terms of both form and sense; (2) such variation is found in different domains, such as the financial and sports domain; (3) existing lexical resources, the Tongyici Cilin in particular as in our current study, should be enriched and enhanced by capturing lexical usages from a variety of Chinese speech communities, to represent the lexical items from a Pan-Chinese perspective; and (4) lexical items obtained from the synchronous Chinese corpus can supplement the existing content of the Tongyici Cilin, with more contemporarily lexicalised concepts, as well as variant expressions of similar and related concepts from various Chinese speech communities.

Hence it remains for us to further investigate how the related lexical items obtained from the synchronous corpus should be grouped and incorporated into the semantic classification of existing lexical resources; and to further explore how they might be extracted in a large scale by automatic means. These will definitely be amongst the most important future directions as discussed in the next section.

[1] Upon further analysis, on average about 60% of these "unique" items were actually found in one or more of the other regions, but with frequency 5 or below. Since the difference in frequency is quite large for most items, we can reasonably treat them as unique to a particular community.

6 Future Work

In the current study, we have investigated the regional variation of lexical items from the financial and sports domain, and the coverage of the Tongyici Cilin for such variation. The results suggested great potential for building a Pan-Chinese lexical resource for Chinese language processing. Our next step would thus be to further investigate more automatic means for extracting the near-synonymous or closely related items from the various subcorpora. To this end, we would explore algorithms like those used in Lin et al. (2003). Of similar importance is the mechanism for grouping the related lexical items and incorporating them into the semantic classifications of existing lexical resources. In this regard we will proceed with further in-depth analysis of the classificatory structures of individual resources and fit in our Pan-Chinese architecture.

Apart from the Tongyici Cilin, there are other existing Chinese lexical resources such as HowNet (Dong and Dong, 2000), SUMO and Chinese WordNet (Huang et al., 2004), as well as other synonym dictionaries from which we might draw reference to build up our Pan-Chinese lexical resource.

7 Conclusion

In this paper, we have investigated the regional variation of lexical items in two specific domains from a synchronous Chinese corpus, and explored their coverage in a Chinese synonym dictionary. Results are encouraging in the sense that 20-40% of the candidate words from various subcorpora are unique to the individual communities, and as much as 70% of such unique items are not yet covered in the Tongyici Cilin. It therefore suggests great importance and potential for a Pan-Chinese lexical resource which we aim to construct. The synchronous corpus is a valuable resource for mining the region-specific expressions while existing synonym dictionaries might provide a ready-made semantic classificatory structure. Our next step would be to explore automatic means for extracting related lexical items from the corpus, and to incorporate them into existing semantic classifications.

Acknowledgements

This work is supported by Competitive Earmarked Research Grant (CERG) of the Research Grants Council of Hong Kong under grant No. CityU1317/03H. The authors would like to thank the anonymous reviewers for comments.

References

Calzolari, N. (1982) Towards the organization of lexical definitions on a database structure. In E. Hajicova (Ed.), *COLING '82 Abstracts*, Charles University, Prague, pp.61-64.

Caraballo, S.A. (1999) Automatic construction of a hypernym-labeled noun hierarchy. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland, pp.120-126.

Chen, K-J., Huang, C-R., Chang, L-P. and Hsu, H-L. (1996) Sinica Corpus: Design Methodology for Balanced Corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC 11)*, Seoul, Korea, pp.167-176.

Dong, Z. and Dong, Q. (2000) *HowNet*. http://www.keenage.com.

Huang, C-R., Chang, R-Y. and Lee, S-B. (2004) *Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO*. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon, Portugal.

Kirkpatrick, B. (1987) *Roget's Thesaurus of English Words and Phrases*. Penguin Books.

Lin, D., Zhao, S., Qin, L. and Zhou, M. (2003) Identifying Synonyms among Distributionally Similar Words. In *Proceedings of the 18th Joint International Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, pp.1492-1493 .

Mei et al. 梅家駒、竺一鳴、高蘊琦、殷鴻翔 (1984) 《同義詞詞林》 (*Tongyici Cilin*). 商務印書館 (Commerical Press) / 上海辭書出版社.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J. (1990) Introduction to WordNet: An online lexical database. *International Journal of Lexicography, 3(4)*:235-244.

Riloff, E. and Shepherd, J. (1999) A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction. *Natural Language Engineering, 5(2)*:147-156.

Sinclair, J. (1987) *Collins COBUILD English Language Dictionary*. London, UK: HarperCollins.

Tsou, B.K. and Kwong, O.Y. (2006) Toward a Pan-Chinese Thesaurus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Tsou, B.K. and Lai, T.B.Y. 鄒嘉彥、黎邦洋 (2003) 漢語共時語料庫與信息開發. In B. Xu, M. Sun and G. Jin 徐波、孫茂松、靳光瑾 (Eds.), 《中文信息處理若干重要問題》 (*Issues in Chinese Language Processing*). 北京：科學出版社, pp.147-165.

Vossen, P., Meijs, W. and den Broeder, M. (1989) Meaning and structure in dictionary definitions. In B. Boguraev and T. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*. Essex, UK: Longman Group.

Xia, F., Palmer, M., Xue, N., Okwrowski, M.E., Kovarik, J., Huang, S., Kroch, T. and Marcus, M. (2000) Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

Region	Unique Items and Examples (Financial)				Unique Items and Examples (Sports)			
HK	**2105 (40.19%)**				**2410 (38.16%)**			
	按揭	證券界	銷情	招股	今季	球證	轉投	客軍
	錄得	開售	地產股	加推	球手	種籽	操練	披甲
	樓盤	收報	寬頻	純利	現時	超聯	友賽	歐國盃
	大市	入市	減價	居屋	港隊	反勝	12 碼	力壓
	息率	新盤	貨尾	低位	決賽周	早前	周一	爭標
BJ	**933 (33.43%)**				**907 (23.39%)**			
	農村	住房	黨	質檢	自行車	攀岩	奧神隊	江蘇隊
	退耕還林	下崗	節水	品種	登山	特級	自行車賽	自治區
	查處	群眾	走私	城鄉	遼寧隊	宏遠隊	前衞	體質
	非典	優化	專項	扶貧	名人戰	中學生	棋院	彩票
	抽查	運行	水資源	林業	山東隊	軍區	散打王	冬訓
TW	**891 (28.85%)**				**1302 (28.30%)**			
	金控	投信	經理人	降息	安打	金剛	牛隊	獅隊
	計劃	釋股	董監事	執行長	中華隊	雷公	三振	主投
	投資人	成長率	團隊	立委	投手	保送	球團	戰神
	網路	升息	坪	個股	職棒	國中	打點	二壘
	營收	買超	契約	專案	全壘打	撞球	鯨隊	大陸隊
SG	**890 (23.20%)**				**1044 (20.39%)**			
	新元	脫售	馬股	私宅	新加坡隊	阿申納隊	芽蘢隊	丹戎巴葛隊
	獻議	港務	財政年	海事	正賽	效勞	大決賽	76 人隊
	閉市	戶頭	文告	財年	新元	利物浦隊	切爾西隊	利茲隊
	公寓	董事部	組屋	辦公樓	射腳	受訪	瓶分	星期六
	平方英尺	共管	地契	輪船	軍團隊	新麒隊	賽項	網隊

Table 6 Uniqueness of individual subcorpora

Region	Financial		Sports	
	Found in Cilin	Not in Cilin	Found in Cilin	Not in Cilin
HK	**560 (26.60%)** 減價 純利 居屋 戶口 拆息 憧憬 容許 倒退 通告 結餘	**1545** **(73.40%)**	**884 (36.68%)** 現時 操練 披甲 晨曦 攻堅 大勇 帝王 蛋 答 爭勝	**1526** **(62.32%)**
BJ	**369 (39.55%)** 農村 抽查 住房 運行 黨 走私 品種 林業 鄉鎮 森林	**564** **(60.45%)**	**355 (39.14%)** 自行車 特級 中學生 前衞 自治區 彩票 農民 解放軍 棋壇 幹部	**552** **(60.86%)**
TW	**265 (29.74%)** 契約 專案 不動產 改選 股利 通路 關卡 週 終場 席次	**626** **(70.26%)**	**354 (27.19%)** 投手 金剛 雷公 保送 打點 地主 水手 報導 總和 晚間	**948** **(72.81%)**
SG	**333 (37.42%)** 公寓 平方英尺 港務 戶頭 共管 文告 地契 海事 輪船 開銷	**557** **(62.58%)**	**281 (26.91%)** 效勞 星期六 星期三 城門 補足 星期四 星期五 星期天 腳踏車 星期二	**763** **(73.08%)**

Table 7 Coverage of the *Tongyici Cilin* for the unique lexical items in individual subcorpora