From Synchronous Corpus to Monitoring Corpus, LIVAC: The Chinese Case

Benjamin K. Tsou Andy C. Chin
Research Centre on Linguistics &
Language Information Sciences
The Hong Kong Institute of Education
Tai Po, Hong Kong
btsou@ied.edu.hk andychin@ied.edu.hk

Oi Yee Kwong
Department of Chinese, Translation & Linguistics
City University of Hong Kong
Kowloon Tong, Hong Kong
olivia.kwong@cityu.edu.hk

***Abstract*—Very large corpora of properly processed textual materials are uncommon but they can provide important resources for language modeling in natural language processing, ranging from speech processing and text input to automatic IR and patent translation. However, when properly cultivated in spatial-temporal terms, they can foster innovative knowledge discovery in database applications by functioning as *monitoring corpus* and enhance the human centered communication environment by allowing more substantive introspection and comparison of linguistic and social-cultural developments of the relevant speech communities.**

This paper discusses how the gigantic synchronous and homothematic corpus of Chinese, LIVAC, can contribute to the monitoring the linguistic homogeneity and heterogeneity diachronically and synchronically. After processing media texts of more than 400 million Chinese characters over 16 years, LIVAC has yielded a lexical corpus of 1.5 million words. This paper examines some aspects of the nature and extent of lexical and morphological divergence and convergence in the Chinese language of Hong Kong, Taipei and Beijing. Additional discussions cover creation and relexification of neologisms, categorial fluidity and the associated challenges to terminology standardization, such as renditions of non-Chinese personal names. This paper also explores how the associated socio-cultural developments can be fruitfully monitored by means of this unique spatial-temporal corpus.

Keywords- monitoring corpus; synchronous corpus; homothematic coprus; LIVAC; the Chinese language

I. INTRODUCTION

Although Chinese is the native or official language in many communities such as Mainland China, Taipei, and Hong Kong, its homogeneity cannot be simplistically assumed. In fact, there are readily noticeable and significant linguistic differences among these Chinese speech communities as any casual newspaper reader from a community other than his or hers will readily testify. This phenomenon can be well illustrated by the lexical items. As a consequence of recent history and localized cultural developments, the differences are arguably much greater than those among British English, American English and Australian English if Chinese-English bilinguals have an opportunity to reflect on the two situations. These linguistic differences are not only significant for NLP and linguistic analysis but for monitoring the speech communities in which these linguistic variations are embedded.

II. USING A SYNCHRONOUS & HOMOTHEMATIC CORPUS FOR MONITORING CHINESE LANGUAGE DEVELOPMENT

In order to explore the significance of the non-homogeneity of the Chinese language in different Chinese speech communities, this paper attempts to exploit a viable and rigorous methodology which can provide, among other things, a useful foundation for research into terminology and standardization of the language.

The use of corpus has been a major means for studying natural language in authentic use rather than in abstraction [1] [2] [3]. There is now an over-abundance of natural language data for constructing linguistic corpora. However, it is important to control the nature, size and time frame of these sources when building corpora, especially when we need to conduct synchronic and/or diachronic comparisons.

Internet has become a major source for obtaining linguistic data because it is easily and readily accessible. However, we have to be cautious when drawing data from the Internet. One commonly seen phenomenon is data duplication where the same data with exact wordings and layout appear more than once on the Internet. Moreover, the timeframes of the data obtained from Internet are neither specified nor easily controlled. Overlooking these problems will lead to serious faults in drawing conclusion, especially when qualitative conclusions are based on the quantitative analysis of these data. Thus it is important to control the data rigorously in terms of both dimensions of time and content.

One major approach in corpus linguistic research is using *balanced corpus* in which data of a language are drawn from a wide range of sources/registers. Examples in the English language include the British National Corpus (BNC) and American National Corpus (ANC). This type of corpus provides a comprehensive overview of the language of a particular community, such as British English and American English. It cannot however compare the same type of language in both spatial and temporal dimensions.

In this paper, we argue that heterogeneity rather than homogeneity should be assumed in the Chinese language, both lexically and syntactically, across some major Chinese communities, such as Beijing, Hong Kong, and Taipei. The LIVAC corpus [4] initially developed at the Language Information Sciences Research Center at the City University of Hong Kong since 1995 is particularly suited for this kind of study. LIVAC is synchronous and homothematic in nature, which rigorously and regularly draws comparable amount of data from similar sections such as front page, financial page, Cross-Strait news page, editorial page, entertainment page, sports page, and local news page, of printed Chinese media of major Chinese communities (see Table I) [5] [6]. In other words, the data are analyzed within the same framework in terms of size, time, domain as well as content across communities and this provides a common platform for meaningful synchronic and/or diachronic comparisons [7]. This "Windows" approach thus ensures that comparable data are extracted according to the same set of criteria [8].

The use of massive news media materials for such a study is very much justified because the popular media should reflect the language and the readership of society and be responsive to their language preference [9] [10] [11] [12]. Moreover, such a database facilitates higher order knowledge discovery and the analysis of associated linguistic characteristics with the larger context of its human users.

Currently LIVAC has obtained 1.5 million word types by accumulatively analyzing over 400 million Chinese characters of newspaper texts in major Chinese communities. [1] Background details of LIVAC are summarized in Table I.

TABLE I. SUMMARY HIGHLIGHTS OF LIVAC CORPUS

Communities covered:	Beijing, Hong Kong, Macau, Shanghai, Singapore, Taiwan, Shenzhen, Zhuhai, Guangzhou
Source of data:	Representative newspapers from each community
Time span:	Since 1995 (i.e., 16 years)
Coverage:	News sections: International, editorials, Cross-Strait, local, financial, entertainment, sports, etc.
Size of corpus:	1.5 million word types culled from 400+ million Chinese character of texts

With the Windows approach described above, the data of Hong Kong, Taipei and Beijing are rigorously processed and are normalized in terms of size and timeframes which can allow us to observe the differential trends of Chinese language and related linguistic developments.

III. LEXICAL CONVERGENCE AND ACTIVE-CORE VOCABULARY

Even though Hong Kong, Taipei and Beijing share the Chinese language, there are significant differences among their large lexical databases. Table II indicates the extent of lexical items shared by the three communities between 1995 and 2004, based on the above Windows approach.

TABLE II. OVERLAPPING LEXICAL ITEMS IN HONG KONG, TAIPEI AND BEIJING BETWEEN 1995 AND 2004

%	Hong Kong	Taipei	Beijing
Hong Kong		39	34
Taipei	41.1		33.6
Beijing	41.1	38.6	

The corresponding percentages between any two communities are not necessarily identical because the total numbers of words in each community are different. Consider Hong Kong and Taipei, 39% of the 215k words from Hong Kong can be found in Taipei while 41.1% of the 205k words in Taipei appear in Hong Kong. These figures show that the number of lexical items actively shared by any two of these three communities over the 9-year period is not very high. Generally speaking, less than half of the lexical items used in any one community can be found in the other two communities. The extent of overlap between Taipei and Beijing is the least. Only 33.6% of such overlapping items are found in the Taipei corpus. This demonstrates that between Taipei and Hong Kong, as well as between Hong Kong and Beijing, there are more extensive overlaps than between Beijing and Taipei. This situation might reflect the social, cultural situation associated with real politics.

In this 9-year period, over 56,000 lexical items were found in common in the **three** communities (see Table III).

TABLE III. OVERLAPPING LEXICAL ITEMS AMONG HONG KONG, TAIPEI AND BEIJING (1995-2004)

No. of overlapping items	Hong Kong	Taipei	Beijing
56693	26.3%	27.7%	31.8%

The extent of pairwise overlap between the communities is given in Figure 1. One may find it surprising that the extent of overlap is exceptionally low even though the three communities share the same language. Table III shows that such a core vocabulary only accounts for from 26% to 31% of the total items used in each community. It should also be noted that the same concept is not necessarily rendered by the same lexical item in the three communities. This is a partial reason leading to the low degree of lexical overlap. Therefore, lexical variation cannot be simply studied on a quantitative basis and a systematic qualitative investigation has to be carried out in order to compare lexical divergence across communities. This will be discussed in Section IV.

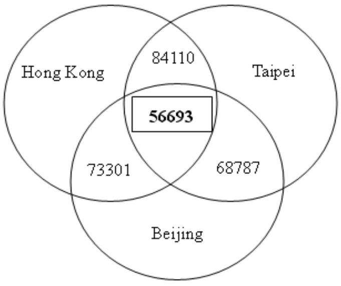

Figure 1. Extent of Lexical Overlap in Hong Kong, Taipei and Beijing in LIVAC (1995 – 2004)

These overlapping words can be considered the **active-core vocabulary** that is in current use in the language. Those non-overlap words can be considered **ambient vocabulary**, which can be divided into two sub-types:

(a) **Transparent and readily decodable**: Even though the three communities use Chinese characters to coin words, mutual intelligibility cannot be always assumed. For

[1] All the texts have been automatically segmented with semi-automatic verification, and large sections have been POS-tagged and verified. Some details on the relevant information mining efforts have been reported in [5] and [6].

example, 計程車 (*ji-cheng-che*, count-distance-car) and 出租車 (*chu-zu-che*, hired-car) are used to refer to "taxi" in Taipei and Beijing respectively, members of each community should be able to understand the other word by adding up the meaning of each morpheme (i.e., 計(count), 程(distance), 出租(hire) and 車(car)). These two words can thus be considered *mutually intelligible* between the two communities, as well as in other Chinese communities such as Hong Kong and Singapore where 的士 (*di-shi*) and 德士 (*de-shi*) are used respectively. Another pair of example is 軟盤 (*ruan-pan*, soft-platter) and 軟碟 (*ruan-die*, soft plate) for *floppy disc*: The former is used in Mainland China and Taipei, while the latter in Hong Kong. 盤 and 碟 are similar in meanings and people should not find problems in understanding the alternate term.

(b) **Opaque and non-readily decodable**: Some words are less mutually intelligible across communities. For example, 的士 (*di-shi*) and 德士 (*de-shi*) meaning "taxi" used in Hong Kong and Singapore respectively cannot be simply derived from the meanings of their components (i.e., 的(*di*), 士(*shi*) and 德(*de*)) because these two words are created by means of phonetic adaptation. They are thus less mutually intelligible to members of Beijing and Taipei.

It is notable that beyond the Chinese context, there is also a considerable degree of mutual intelligibility between words used in Chinese and Japanese (i.e., those written with Japanese kanji) on vehicle-related words. Chinese readers are found to be able to understand more vehicle-related words written with Japanese kanji than vice versa [13]. It is also noted that the extent of overall mutual intelligibility for understanding Chinese items by Japanese has decreased from 51% to 25% when the same window is taken 10 years later [14] and this deserves fuller investigation. One possible reason for the decrease is that many VEHICLE words in Chinese are phonetically adapted (mostly from English), such as *ji-pu-che* 吉普車 (jeep), *mo-tuo* 摩托 (motorbike). The meanings of these words are not transparent from the Chinese characters, i.e., the meaning of 摩托 is not simply a combination of the meanings of 摩 and 托.

Although 的士 (*di-shi*), 計程車 (*ji-cheng-che*, count-distance-car) and 出租車 (*chu-zu-che*, hired-car) appear in all three communities, their frequency distributions are significantly different across the three communities: 的士, 計程車 and 出租車 are predominantly used in Hong Kong, Taipei and Beijing respectively. It is thus more appropriate to consider the other two items ambient vocabulary in each community. In this regard, we re-define *ambient vocabulary* as those items whose frequency in a particular community accounts for 80% or above of the total frequencies from all the three communities. Table IV provides the quantitative data of this type of vocabulary in the three communities.

TABLE IV. WORDS WITH OVER 80% LOCAL USAGE FREQUENCY

%	**Hong Kong**	**Beijing**	**Taipei**
Type	55.8	51.8	59.5
Token	12.8	9.9	11.7

The data show that about half of the lexical items of each community have high local usage frequency. In terms of tokens, these local high frequency items only account for around 10% of the overall token usage. This again demonstrates that there is a high degree of lexical heterogeneity across these three Chinese communities.

The above discussion draws attention to the considerable heterogeneity of the Chinese language among Hong Kong, Taipei and Beijing. Furthermore, we should also point out that non-reciprocal items found in one single community will subsequently spread to other communities upon frequent cross-communal contact and will gradually become the active-core vocabulary. This dynamic nature of lexical development will be explored in the next section.

IV. RENDITIONS OF FOREIGN PERSONAL NAMES

The tremendous growth and attrition of proper names in the Chinese language has become a challenge in NLP, especially for named-entity recognition [15] [16]. Chinese, unlike English, does not have any means, such as capital letters to identify proper names. Thus, in LIVAC, three types of proper names (personal names, geographical names and organization names) are separately tagged in multiple ways.

While phonetic adaptation is commonly used to render foreign personal names in Chinese, the three communities show considerable variations which are critical to NLP in a cross-linguistic context. Such differences can be attributed to the use of different dialects for the transliteration template. For example, Cantonese is the local dialect providing the usual base for phoneticization in Hong Kong while Mandarin is the base for Beijing and Taipei. Furthermore, even for those popular figures whose names appear frequently in news media, discrepancies across communities also exist so that members from one community may not recognize readily that two different Chinese renditions in fact could refer to the same individual [17] [18]. Table V lists some well-known non-Chinese names rendered differently in the three communities, according to LIVAC.

TABLE V. NON-CHINESE PERSONAL NAMES WITH DIFFERENT RENDITIONS IN LIVAC

Names	**Hong Kong**	**Taipei**	**Beijing**
George W. Bush	布殊	布希	布什
Tony Blair	貝理雅	布萊爾	布萊爾
Saddam Hussein	薩達姆	哈珊	薩達姆
Zinedine Zidane	施丹	席丹	齊達內
Whoopi Goldberg	胡比高拔	琥碧戈柏	烏比·戈德堡
Brad Pitt	畢彼特	布萊德彼特	布拉德皮特

Besides the dialects involved in the recipient language, there are other communal differences, such as the number of syllables in the transliteration, as shown by the renditions of Brad Pitt and Whoopi Goldberg. Furthermore, even within the same community, different domains might have different principles for transliteration. For example, in the domain of entertainment, both first name and last name are always included in the transliteration, while in the political and sports domains, only the last names are transliterated.

V. Relexification

The lexical divergence across Chinese communities can be reduced through **relexification** [8]. In the initial stage, there can be alternate lexical items referring to the same concept in different communities. Subsequently, these lexical variants compete among each other and some are retained and become core items. *Internet* and *mobile phone* are good examples to illustrate the relexification process.

A. Internet

The rapid developments in computer technology have led to the coinage of new words. The lexical variation in the IT domain can be best illustrated by those words designating *Internet*. In LIVAC, there have been at least 13 lexical items referring to this technology since it was first introduced, as shown in Table VI below.

TABLE VI. Alternate Renditions of "Internet" in LIVAC

1. 互聯網 mutual-link-net	8. 網際網絡 inter-net-network
2. 互聯網絡 mutual-link-network	9. 網際網路 inter-net-network
3. 交互網 cross-mutual-net	10. 遞訊網 transmit-information-net
4. 信息網 information-net	11. 英特網 INTER-net
5. 訊息網 information-net	12. 因特網 INTER-net
6. 國際網 international-net	13. 萬維網 10K-dimension-net
7. 國際聯網 international-link-net	

The data show that when Internet was first introduced, there were diverse renditions for this technology in the Chinese communities. Items 1 – 10 are created by means of semantic adaptation with the functions and characteristics of Internet being described. Items 11 - 13 are created by means of phonetic adaptation or hybrid (a combination of both semantic and phonetic adaptations) by which the pronunciation of "internet" in English is modeled. It is interesting to observe that after subsequent relexification and merger, 互聯網 (mutual-link-net) became the most popular term with 因特網 (INTER-net) as the next frequently used item by year 2000 (for more details on Chinese neologistic development, see [7], [19] and [20]).

B. Mobile phone

The LIVAC data point to at least 10 items referring to *mobile phone* in Chinese, as shown in Table VII below:

TABLE VII. Alternate Renditions of "Mobile Phone" in LIVAC

手持電話 hand-hold-phone	無線電話 no-wire-phone
手提電話 hand-carry-phone	隨身電話 follow-body-phone
行動電話 action-phone	攜帶電話 carry-phone
流動電話 transient-phone	大哥大 Big-Boss-Brother
移動電話 mobile-phone	手機 Hand-phone

We find significant convergent developments in Hong Kong, Taipei and Beijing, and discrete changes in the choice among alternate forms when comparison is made on with three consecutive annual windows (from 1998 to 2001), as shown in Table VIII below:

TABLE VIII. Development of Lexical Items Related to "Mobile Phone" from 1998 to 2001

Years	Hong Kong	Taipei	Beijing
98-99	手提電話 hand-carry-phone	行動電話 action-phone	移動電話 mobile-phone
	流動電話 transit-phone	大哥大 Big-Boss-Brother	大哥大 Big-Boss-Brother
99-00	流動電話 transit-phone	行動電話 action-phone	手機 hand-phone
	手機 hand-phone	手機 hand-phone	移動電話 mobile-phone
00-01	流動電話 transit-phone 手機 hand-phone	手機 hand-phone	手機 hand-phone
	--	行動電話 action-phone	移動電話 mobile-phone

The three communities initially had different neologistic renditions for *mobile phone*. 手提電話 (*shou-ti-dian-hua*, hand-carry-phone), 行動電話 (*xing-dong-dian-hua*, action-phone) and 移動電話 (*yi-dong-dian-hua*, mobile-phone) were used most frequently in Hong Kong, Taipei and Beijing respectively. In 1999-2000, the disyllabic item 手機 (*shou-ji*, hand-phone) was the most frequently used in Beijing while it was the next frequently used item in Hong Kong and Taipei. In 2000-2001, it completely took over other items and became the core item for *mobile phone* in all three communities. There can be a number of reasons for one item winning over the others and disyllabification could be one of such reasons since it is the major trend of lexical development in the Chinese language. According to Masini's study, the ratio between monosyllabic and poly-syllabic words is approximately 1:6. Among these polysyllabic words, over 70% are disyllabic [21]. In her sociolinguistic study on the nature of "Chinese word" [22], Wang found that over 90% of the words segmented by her informants are disyllabic. The propensity for disyllabification has often been noted [23] [24] [25] [26].

VI. Categorial Fluidity in Chinese

Chinese is an isolating language which lacks morphological markings to distinguish different parts-of-speech (POS). For example, the word 懷疑 (*huai-yi*), with a verb sense ("to suspect") to start with, should only appear as a verb in a dictionary, despite its variable usages found in real contexts, as in (a) – (c) below.

(a) 我 懷疑 他 是 賊
wo huai-yi ta shi zei
'I suspect he is a thief'

(b) 他 滿臉 懷疑 表情
ta man-lian huai-yi biao-qing
'He wears a suspicious look'

(c) 這 只 是 我 的 懷疑
zhe zhi shi wo de huai-yi
'This is only my suspicion'

In (a), 懷疑 (*huai-yi*) is a verb. In (b), it is an adjective and in (c), it is a noun.

We call this relative flexibility of a word being used for different grammatical functions and possibly different POSs *categorial fluidity* [27]. In the following, we only focus on the fluidity between verbs and nouns. We consider categorial fluidity a continuum. We will show, by means of the LIVAC data, how categorial shift takes places across Chinese communities and over time.

A. *Methodology*

First, all verbs (excluding copula verbs and auxiliary verbs) with their frequencies were extracted. Then out of these verbs, those which also exhibited noun or nominalized usages were extracted together with the corresponding frequencies. We call this set of verbs "VN words".

Next, a simple ratio (1) was computed for all VN words. The log ratio was used to give a linear scale. If verb usage outnumbers noun usage to a certain extent, i.e., when $r >> 0$, it suggests that the word is originally a verb and has just started to shift. If verb usage and noun usage are more or less equal, i.e., when $r \approx 0$, then either the shift is mature enough or there is genuine ambiguity. If noun usage outnumbers verb usage by a lot, i.e., when $r << 0$, it would mean that either the verb has over shifted or the word is originally a noun and is occasionally denominalized.

$$r = \log_2 \frac{verb\ uses}{noun\ uses} \quad (1)$$

B. *Results*

The results from Hong Kong, Beijing and Taipei are shown in Table IX. The column "No. in VN Shift" indicates the amount of VN words out of all verbs. Out of these we analyzed those with total frequency (including. both verb and noun usages) 5 or more for their r values, and the results are shown in the last three columns. Each place has about 60% of the words reaching this threshold. With $r \geq 1$, verb usage at least doubles noun usage. With $1 > r > -1$, verb and noun usages are quite balanced. With $r \leq -1$, noun usage at least doubles verb usage. The results thus suggest that in real use, there are about 3-4% more nominalized uses of verbs found in Beijing than in Hong Kong and Taipei, which indicate quite a different, if not innovative, style of writing in Beijing. The figures also reflect the asymmetry between deverbalization of verbs and denominalization of nouns.

TABLE IX. SUMMARY OF RESULTS (HONG KONG, BEIJING & TAIPEI)

Source	No. of VN	$r \geq 1$	$1 > r > -1$	$r \leq -1$
		(Only for word types with freq≥5)		
Hong Kong	14.4%	51.6%	26.0%	22.4%
Beijing	18.5%	45.5%	29.7%	24.8%
Taipei	15.5%	49.1%	27.1%	23.8%

The general observation is that Beijing demonstrates more nominalized usages of verbs than the other two communities. In addition, for Hong Kong and Taipei, on average about 50% of the VN words are just beginning to shift (with $r \geq 1$) and their verb usages are still dominant. On the contrary, only about 35% on average of the VN words have $r \geq 1$ for Beijing. In other words, many words which are verbs originally do not actually play the role of verbs in Beijing. This might suggest that Chinese grammar is more seriously Europeanized in the Mainland.

We also found that 105 VN words are shared by all three communities and their usages are quite different among the three communities. Some examples and the corresponding r values are shown in Table X.

TABLE X. EXAMPLES OF VN WORDS COMMON TO THREE PLACES

VN Word	Hong Kong	Beijing	Taipei
發揮 (to express)	0.8074	2.7004	3.3219
經營 (to run a business)	1.5850	-0.1375	1.8074
宣傳 (to promote)	-1.3785	1.5850	0.4854
合作 (to co-operate)	1.3785	-1.1635	0.4594
衝擊 (to attack)	-0.3219	1.3219	-3.3219
感受 (to experience)	2.8074	-1.4150	-1.3219

Table X shows that 發揮 (*fa-hui*, to express) maintains most of its verb usage in Beijing and Taipei, but is considerably balanced with its noun usage in Hong Kong. On the other hand, 感受 (*gan-shou*, to experience) is mostly used as a verb in Hong Kong, but its noun usage predominates in Beijing and Taipei.

The above comparison indicates that the categorial fluidity phenomenon is relatively most common in Beijing – up to more than 18% of verbs undergo the verb-noun transitional process to various extents. These findings not only improve our understanding of this perennial problem in contemporary Chinese, but also have important implications for meaningful natural language applications.

VII. SYNTACTIC CHANGE

Besides lexical development, LIVAC also allows us to monitor syntactic change of the Chinese language. With its synchronic nature, we can trace how the new syntactic feature originates. In the following, we discuss the transitive verb 打造 (*da-zao*, to fabricate).

In the *Dictionary on Modern Chinese* published in 2003, the verb 打造 has the following definition:

"to fabricate (mostly metallic objects such as tools and ships)"

In a later edition published in 2005, one more definition has been added: "to create or to accomplish something such as brand names or company image".

These two definitions show that the objects of 打造 change from *concrete* to *less concrete* or even *abstract*. It is thus meaningful to trace how this property of the syntactic argument changed. To be more specific, can we trace which Chinese community instigates this change first?

In LIVAC, the objects of 打造 are classified into 3 types:

(a) Concrete objects, such as ships, furniture

(b) Semi-concrete objects such as the aircraft carrier of the automotive business.

(c) Abstract objects, such as New Taiwan, Peking operas, new brand names, new vista, new life, etc.

When we compare the change of the syntactic object in terms of abstractness for 打造 across the three communities, we find that Taipei was the first to take on *abstract* objects for this verb, followed by Hong Kong and Singapore, then Shanghai and Beijing (non-government publications). The chronological order is summarized in Figure 2.

Taipei (May 1997)
⇩
Hong Kong (July 1998), Singapore (Sept 1998)
⇩
Shanghai (Sept 1999)
⇩
Beijing A (non-government publications)
⇩
Beijing B (People's Daily, Aug 2000)

Figure 2. Chronological development of the abstractness of the syntactic object for 打造 (to fabricate) across Chinese communities[2]

VIII. Concluding Remarks

In this paper, we have drawn attention to the innovative use of a gigantic corpus of Chinese which has been cultivated synchronously and homothematically and with the introduction of a Windows approach. Such a linguistic corpus provides much more value than for traditional language modeling efforts in NLP applications such as IR and named-entity recognition. It can function usefully for data mining as well as monitoring linguistic variations in spatial and temporal dimensions which is uncommon for traditionally morphology rich languages, as well as to monitor the deeper concomitant developments in the larger relevant social and cultural contexts with the associated language users. It is hoped that with the addition of more mature applications of data mining techniques, much more findings can be reported in future.

Acknowledgment

We have benefited from input by many colleagues: W. F. Tsoi, K. P. Chow, Tom Lai, Terence Chan, Amy Liu, and colleagues from the ChiLin Star Corporation in Zhuhai, PRC.

References

[1] K. Church and R. Mercer, "Introduction to the Special Issue on Computational Linguistics Using Large Corpora," Computational Linguistics, vol. 19.1, 1993, pp. 1-24.

[2] T. McEnery and A. Wilson, Corpus Linguistics. Edinburgh: Edinburgh University Press, 1996.

[3] D. Biber, S. Conrad, and R. Reppen, Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998.

[4] http://www.livac.org

[5] B. Tsou, H. L. Lin, T. Chan, J. Hu, C. H. Chew, and J. K. P. Tse, "A Synchronous Chinese Language Corpus from Different Speech Communities: Construction and Applciation," International Journal of Computational Lingusitics and Chinese Language Processing, vol. 2.1, 1997, pp. 91-104.

[6] B. Tsou, A Synchronous Dictionary on Pan-Chinese Syntactic Information, unpublished [in Chinese].

[7] B. Tsou and T. B. Y. Lai, "Chinese synchronous corpus and information mining," in Critical Issues in Chinese Information Processing, B. Xu, M. Sun and G. Jin, Eds. Beijing: Science Press, 2003, pp. 147-165 [in Chinese].

[8] B. Tsou, "A window on re-lexification in Chinese," in In Memory of Professor Li Fang-kuei: Essays on Linguistic Change and the Chinese Dialects, P. H. Ting and A. Yue, Eds. Seattle/Taipei: University of Washington/Academia Sinica, 2000, pp. 53-72.

[9] T. A. van Dijk, News Analysis: Case Studies of International and National News in the Press. Hillsdale: Lawrence Erlbaum, 1998.

[10] T. A. van Dijk, News as Discourse. Hillsdale: Lawrence Erlbaum, 1998.

[11] P. Garrett and A. Bell, "Media discourse: A critical overview," in Approaches to Media Discourse, A. Bell and P. Garrett, Eds. Oxford: Blackwell, 1998, pp. 1-20.

[12] R. Fowler, Language in the News: Discourse and Ideology in the Press. London: Routledge, 1991.

[13] B. Tsou and L. Feng, "A comparative study on neologisms in Chinese and Japanese: Towards a windows approach on the creation of neologisms and relexification," Studies on Language, vol. 3, 2000, pp. 51-70 [in Chinese].

[14] B. Tsou and A. Chin, "A large synchronous corpus as monitoring corpus: Some comparative content analysis of Chinese and Japanese language developments," in Proceedings of the 4th International Universal Communication Symposium (IUCS 2010), IEEE Computer Society, in press.

[15] M. Sun, H. Huang, and J. Fang, "Identifying Chinese names in unrestricted texts," Journal of Chinese Information Processing, vol. 9.2, 1998, pp. 16-27 .

[16] L. Cheung and B. Tsou, "Personal names in unrestricted Chinese texts: nature and identification," in Proceedings of Workshop on International Standards of Terminology and Language Resource Management, Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, May 28 – June 2, 2002, pp. 1-7.

[17] B. Tsou and O. Kwong, "Aspects of MT requirements related to proper nouns in the Asian context" in Proceedings of Workshop on Survey on Research and Development of Machine Translation in Asia, 2002, pp. 85-93.

[18] R. Song and B. Tsou, "A preliminary study on Chinese proper names," in Proceedings of the 20th Anniversary Conference of CIPSC, 2000, pp. 14-19. [in Chinese].

[19] B. Tsou and R. J. You, A Dictionary of Chinese New Words in the 21st Century, Shanghai: Fudan University Press, 2007. [in Chinese]

[20] B. Tsou and R. J. You, A Dictionary of Chinese New Words, Beijing: Commercial Press, 2010. [in Chinese]

[21] F. Masini, The Formation of Modern Chinese Lexicon and its Evolution toward a National Language: The Period from 1840 to 1898. Journal of Chinese Linguistics, Monograph 6, Berkeley: Journal of Chinese Linguistics, 1993.

[22] L. Wang, A Sociolinguistic Study on Chinese Words, Beijing: Commercial Press, 2003. [in Chinese].

[23] W. Pan, B. Ye, and Y. Han, A Study on Word Formation in Chinese, Taipei: Students Publisher, 1993. [in Chinese].

[24] J. Zhou, The Meaning and Structure of Words, Tianjin: Tianjin guji chubanshe. 1994. [in Chinese].

[25] S. Lü, "A preliminary study on disyllabicity in modern Chinese," in Collection of Essays on Chinese Grammar, Beijing: Commercial Press, 1999, pp. 415-444. [in Chinese].

[26] Z. Tang, The Lexicon of Contemporary Chinese: Its Synchronic Situation and Change, Shanghai: Fudan University Press, 2001. [in Chinese].

[27] O. Kwong and B. Tsou, "A synchronous corpus-based study of verb-noun fluidity in Chinese," in Proceedings of the 17th Pacific Asia Conference, October 2003, pp. 194-203.

[28] B. Tsou, A. Chin, and O. Kwong, "On incipient linguistic variations in chinese: A corpus approach," unpublished.

[2] A more detailed study on monitoring syntactic change in the Chinese language with the LIVAC corpus can be found in [28].