Computational Linguistics and Chinese Language Processing
vol.2, no.1, February 1997, pp. 91-104.

A Synchronous Chinese Language Corpus from Different Speech Communities: Construction and Applications

Benjamin K. T'sou*, Hing-Lung Lin*, Godfrey Liu*, Terence Chan*
Jerome Hu, Ching-hai Chew+, John K.P. Tse++**

Abstract

Similar to other languages such as English, Spanish and Arabic, Chinese is used by a large number of speakers in distinct speech communities which, despite sharing the unity of language, vary in interesting ways, and a systematic study of such linguistic variation is invaluable to appreciate the diversity and richness of the underlying cultures. This paper describes Project LIVAC (Linguistic Variation in Chinese Communities), which focuses on the development of a Chinese corpus, based on data taken concurrently at regular intervals from multiple Chinese speech communities. The resulting database and computerized concordance from the approximately 20 million word corpus with uniform time reference points extending across two years enable linguists and social scientists to undertake meaningful qualitative and quantitative comparative analysis of the development of linguistic and cultural variation. To facilitate these studies, a framework for integrating the corpus with specific corpus analysis applications is proposed. Based on this framework, a prototype retrieval system, which supports longitudinal studies on word and concept distribution, as well as lexical and other linguistic variation, is designed and implemented.

1. Introduction

The Chinese language, with the largest number of native speakers and users in the world, is generally assumed to enjoy a high degree of uniformity, even though the users of the language recognize important internal variations. Comparative studies of language variation in different Chinese communities are invaluable for the specific investigation of the increasingly apparent language innovation and differential longitudinal development regarding the use of the Chinese language in major Chinese speech communities, and for the study of language variation in space and time in general. In the case of

* Language Information Sciences Research Centre, City University of Hong Kong, Hong Kong.
E-mail: adbigben@cityu.edu.hk

+ Chinese Language Research Centre, Nanyang Technological University, Singapore.

++Department of English, Taiwan Normal University, Taiwan.

**Department of Linguistics, Lingnan College, Hong Kong.

Mandarin based written Chinese, it has been demonstrated that the use of on-line corpora in linguistic studies allows linguists to access information about its lexical or structural features in order to investigate their distribution properties [Chen *et al.* 1993]. To undertake a systematic comparative study of linguistic norms and variation in different Chinese speech communities, it is necessary to build up a synchronous Chinese corpus that goes beyond traditional collection and taxonomy of lexical items from one single community. It is important to support more specific efforts such as:

1. variation studies on where and how a linguistic innovation spreads to adjacent areas [T'sou 1989];
2. studies on differential longitudinal developments in different communities [Chen 1984, Tse 1986, T'sou 1989];
3. lexical studies involving importation and substitution of new lexical elements and semantic relexification [T'sou *et al.* 1990] in specific domains, such as science and technology, business and finance, law and politics, or in the language as a whole;
4. other research endeavours in social sciences, such as sociology, political science and mass communication, based on content analysis of synchronous and homeothematic data [T'sou *et al.* 1990].

To carry out the above mentioned linguistic studies, a large amount of textual material must be collected from various sources in different Chinese speech communities to form a text corpus. While several Chinese text corpora have been available for consultation, each contains only data pertaining to a single specific Chinese speech community, such as Mainland China [17, 18], Taiwan [CKIP 1993] or Hong Kong [16]. There was no single synchronized text corpus from different Chinese speech communities until the present corpus described here came into existence.

It may be noted that while one text corpus can support many different kinds of linguistic studies (called corpus applications), different forms of word concordance and different procedures for corpus analysis must be developed for each different corpus application. To facilitate the development of such corpus applications, we propose a framework for integrating the text corpus with a specific corpus application to generate a word concordance, which can be queried by a linguist or social scientist in conducting research.

2. A Framework for Corpus Application Development

Our framework, as shown in Fig. 1, is based on the process of data modeling [Sanders 1995], which has been well developed in the field of database design. For a given corpus

application, the responsible linguist specifies the data requirements of the related linguistic study. The application developer formalizes these data requirements with the conceptual modeling tool of the entity relationship (ER) diagram [Chen 1976]. The application developer may or may not be the same person as the linguist. The ER diagram is then mapped to a relational data model, which can be easily implemented using an existing Database Management System (DBMS).

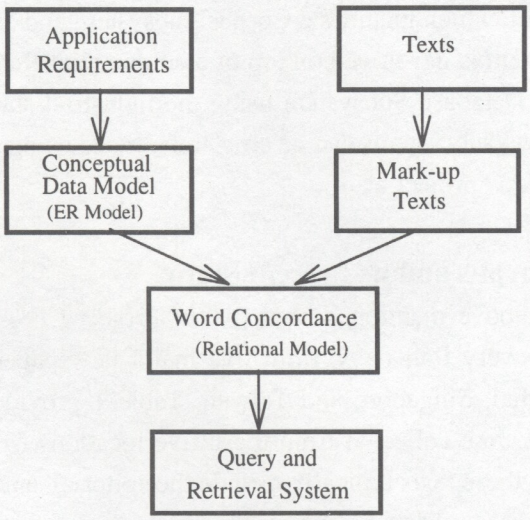

Figure 1 *A Framework for Corpus Application Development*

The original text corpus is transformed by marking up text strings in the corpus according to some predetermined linguistic criteria. Depending on the nature of the corpus application, this mark-up process can be carried out by a computer or by human operators. Furthermore, a mark-up language is used to facilitate the transformation between the source text structure and the specific data structure of the word concordance, and to automate the linkage between these two information bases.

Finally, the mark-up text is processed according to the conceptual data model to produce a word concordance, in the form of a relational database, for the specific application. This relational database, together with the capabilities provided by the mark-up language processor, can be queried by end-users for the purpose of the intended research.

3. Description of the LIVAC Corpus and its Database

LIVAC (for LInguistic VAriation in Chinese communities), which began in 1993, is a major linguistic research project undertaken by the Language Information Sciences Research Centre (LISRC) at the City University of Hong Kong. One major objective of

this project is to construct, over an extended period, a unique 20 million character corpus drawn from the language of different Chinese speech communities, coordinated with uniform calendar reference points. Additionally, it will be developed into an integrated, user-friendly environment for knowledge exploration of the constructed corpus by linguists and social scientists.

In order to achieve the above objective, the LIVAC System is designed as an integrated environment, which includes a Corpus Subsystem and a Database Subsystem. The Database Subsystem is developed on top of a commercial Relational DBMS. Users can interact with the Database Subsystem using the industrial-standard database query language SQL. The two subsystems can be cross-referenced using custom-made utilities developed by the LIVAC project team.

3.1 The LIVAC Corpus and its Present Status

In order to meet the above mentioned objective of Project LIVAC, data are collected synchronously, once every four days, from five major newspapers published in Hong Kong, Macau, Shanghai, Singapore and Taiwan. Table 1 provides details of the synchronous texts which were collected from these five locations from July 1995 to June 1996. The contents of these texts typically include the editorial, and all the articles on the front page, international and local news pages, as well as features and reviews. The quantity of the text data collected per newspaper per day is targeted to be around twenty thousand Chinese characters, so sometimes minor feature stories have been omitted when the data available has exceeded this target. But there have also been instances where this target has not been reached.

Table 1. *Data collected synchronously from July 1995 to June 1996 for five Chinese speech communities*

Region	Newspaper	No. of Days
Hong Kong	明　　報	91
Macau	華 僑 報	91
Taiwan	中央日報	90
Shanghai	新民晚報	91
Singapore	聯合早報	90
	Total	453

Total Number of Characters: 8.3 million

Data collection from multiple locations is a non-trivial task. In the early stages of this project, data, in the form of electronic text files, were delivered to LISRC by email via the Internet, or by regular mail. This not only caused a time delay, but has also resulted in a loss of data as can be seen in Table 1. More recently, because of the rapid development of the World-Wide-Web (WWW), we have been able to down-load many of our needed newspaper articles from the Web.

Upon receipt of the initial data, the tasks of code unification (involving the Big-5 code and the GB code), text formatting and word segmentation are conducted to produce a corpus of verified text. Code unification is needed to convert all the texts in GB code into the standard Big-5 code for further processing. Because conversion from GB code to Big-5 code is a one-to-many mapping, human operators are needed to correct errors at the word level after the computerized code conversion. This latter operation by human operators is further assisted by in-house software developed by the programming staff of Project LIVAC.

Because there is more than one standard for segmenting Chinese text, our segmentation process for the verified text takes as a starting point the national standard GB13715 published in the PRC. Extensive enhancement of the GB13715 standard is necessary because GB13715 is syntax-based and is designed to be applicable to segmentation of common words while words of particular interest to LIVAC are usually conceptual or semantic-based. New segmentation rules were needed to satisfy the requirements of LIVAC, and the word segmentation rules used in LIVAC will be published separately. Furthermore, in order to accommodate users who might prefer a different segmentation standard, the texts in our corpus exist in two versions: the original version (without segmentation) and the verified version (after segmentation according to the enhanced GB13715 or GB13715e [LIVAC 1997]).

3.2 The LIVAC Database

The major research problems that motivate the development of the LIVAC corpus are longitudinal and comparative studies on the quantitative and qualitative aspects of word frequency distribution and developments in lexical variation of modern Chinese. We are specifically interested in those words which have come into the Chinese language since the major language reform efforts of the May Fourth Movement of 1919. We call this set of words New Concept Words (abbreviated as NC Words below), in contrast to other Chinese words that may be called common words.

In order to study (1) the structure and formation of these NC Words, (2) the origin of lexical innovation, and (3) the spread of NC Words to adjacent regions, a classification scheme is introduced to capture the essential features of an NC Word, which are useful

for these research purposes.

This classification scheme, when applied to an NC Word, consists of three orthogonal dimensions as follows:

1. What is the source of the NC Word? The possible sources include foreign languages, classical Chinese, dialects, and new formation. Words from each source are further classified into different types according to the ways in which they are formed. Details of this classification criterion are shown in Fig. 2.
2. Is the NC Word abbreviated?
3. What object class does the NC Word belong to? This criterion is reserved for researchers who are interested in specific classes of NC Words, e.g., person names, places, organizations etc.

An ER diagram is employed to specify what features are relevant to the study of lexical variation [T'sou 1995, LIVAC 1997] of the set of NC Words. We have identified three basic entities, whose attributes are defined below:

NEWS(MEDIA,NEWS_ORG,DATE,SERIAL#,HEADLINES,REGION,INT_CODE)
WORD (WORD,FREQ_COUNT)
NC_WORD (NC_WORD,PINYIN,ENGLISH_EQ,TYPE,ABBREVIATION,OBJECT)

Fig. 3 shows the entity relationship diagram (attributes excluded) for this application, and the meanings of each entity type are explained below.

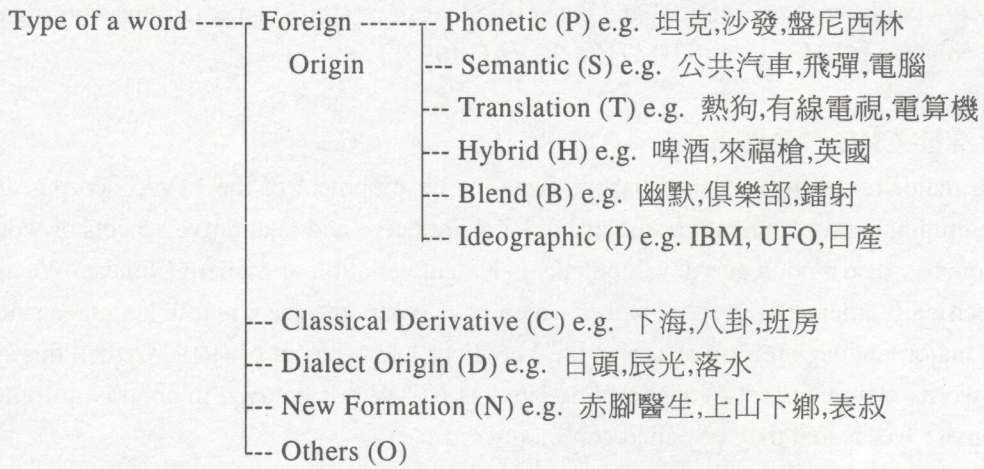

Figure 2 *Classification of an NC Word according to its source and formation rule*

Each NEWS entity identifies a piece of news stored in the corpus. This piece of news may be a newspaper article or, additionally, a television or a radio script. This information is stored in the attribute MEDIA. Where and when this piece of news was published or broadcast are stored in the attributes NEWS_ORG and DATE, respectively. SERIAL# is used to distinguish between different pieces of news from the same news organization on the same date. The title of the news item is stored in the attribute TITLE while the speech community in which this piece of news was written is kept in the attribute REGION. Since these electronic news files are collected from different sources using different internal codes (e.g. GB, Big-5, etc.), it is necessary to retain information of their internal codes in order to allow different codes to be unified before further processing takes place, or to be traced back to the original words if necessary. The attribute INT_CODE is included for this purpose. It should be noted that GB code is usually associated with text printed in simplified Chinese characters while Big-5 code is associated with traditional Chinese characters.

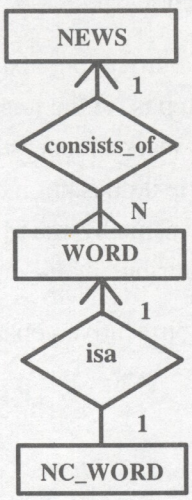

Figure 3 *The ER Diagram for the NC Word Application*

Each instance of the NEWS entity consists of a series of words. Each word is made up of one or more Chinese characters. The word frequencies from different regions are produced for comparative study of relative word usage as a function of different Chinese

speech communities. The WORD entity is used to record each distinct word (with the attribute WORD) and its frequency count (with the attribute FREQ_COUNT) with respect to an instance of NEWS.

The entity NC_WORD is a subclass of the entity WORD. This entity has attributes regarding the properties of an NC Word, including its pinyin romanization (with the attribute PINYIN) and its English equivalent (with the attribute ENGLISH_EQ). An NC Word is classified according to the three classification criteria discussed above, namely, the type of NC Word (with the attribute TYPE), its abbreviation (with the attribute ABBREVIATION), and its object class (with the attribute OBJECT). The NC_WORD entity will be the basis for conducting sociolinguistic study of lexical importation, innovation and variation in the Chinese language [T'sou 1975, 1993].

3.3 From the LIVAC Corpus to the LIVAC Database

News data are collected and stored in their original form and in a segmented form within the LIVAC Corpus, but the information needed by the LIVAC Database must be extracted from each text in the LIVAC Corpus to fill up the attribute values of the LIVAC Database. It is necessary to set up a link to automate data extraction, transformation and reference between these two information data bases.

After the tasks of internal code unification, text formatting, and word segmentation are completed, the next processing step is for the linguist to directly mark up each text on the screen. The main purpose of this mark-up operation is to identify the NC Words in a text. Furthermore, since the original text does not include any information about the text structure, additional markers will be manually added to facilitate automated extraction of information from the text to fill the attribute values of the database.

Figures 4-6 below show three consecutive versions of the same news article that are available from the LIVAC corpus.

埃及蘇丹兩國關係惡化 @

隨著與鄰國埃及的關係日益緊張，蘇丹政府威脅要切斷流往埃及的尼羅河水源，然而對於這個說法，一名國際水利專家卻認爲是天方夜談，絕無可能在不久將來實現。由於埃及政府指摘蘇丹日前策劃行刺其總統穆巴拉克，令到雙方關係急轉直下。

Figure 4 *A news article in its original form*

< 埃及 > < 蘇丹 > < 兩 > < 國 > < 關係 > < 惡化 > @
< 隨著 >< 與 >< 鄰 >< 國 >< 埃及 >< 的 >< 關係 >< 日益 >< 緊張 > ， < 蘇丹 >
< 政府 >< 威脅 >< 要 >< 切斷 > < 流 >< 往 >< 埃及 >< 的 >< 尼羅河 >< 水源 > ，
< 然而 > < 對於 > < 這個 > < 說法 > ， < 一 >< 名 >< 國際 >< 水利 >< 專家 > < 卻 >
< 認爲 >< 是 >< 天方夜談 >< 絕 >< 無 >< 可能 >< 在 >< 不久 >< 將來 >< 實現 > 。
< 由於 >< 埃及 >< 政府 >< 指摘 >< 蘇丹 >< 日前 >< 策劃 > < 行刺 > < 其 >< 總統 >
< 穆巴拉克 > ， < 令到 >< 雙方 >< 關係 >< 急轉直下 > 。

Figure 5 *The news article of figure 4 after word segmentation*

< 埃及 px> < 蘇丹 px> < 兩 > < 國 > < 關係 > < 惡化 > @
< 隨著 > < 與 > < 鄰 > < 國 > < 埃及 px> < 的 > < 關係 > < 日益 > < 緊張 > ，
< 蘇丹政府 h> < 威脅 > < 要 > < 切斷 > < 流 > < 往 > < 埃及 px> < 的 > < 尼羅河 h>
< 水源 > ， < 然而 >< 對於 >< 這個 >< 說法 > ， < 一 > < 名 > < 國際水利專家 s >
< 卻 >< 認爲 >< 是 >< 天方夜談 s> ， < 絕 > < 無 > < 可能 > < 在 >< 不久 >< 將來 >
< 實現 > 。 < 由於 > < 埃及政府 h > < 指摘 > < 蘇丹 px > < 日前 > < 策劃 > < 行刺 >
< 其 > < 總統 s> < 穆巴拉克 pn> ， < 令到 > < 雙方 > < 關係 > < 急轉直下 > 。

Figure 6 *The news article of figure 5 after mark-up*

After the mark-up operation is completed, the resulting text is processed by a computer program, which scans the tagged text to extract information to fill the attribute values of the database directly or to perform additional computations (e.g., frequency count) for derived attribute values. Note that the classification and the English equivalent (i.e., meaning) of each new entry of an NC Word must be explicitly assigned by a linguist. We have found that this is the most costly and labor-intensive operation of the whole project. The resulting database can be queried by a linguist for various purposes related to the problem of lexical variation in Chinese.

4. Content Analysis of Synchronous and Homeothematic Data in LIVAC

Because of the synchronous nature of the data, LIVAC provides a unique means of making a systematic cross-community comparison of written Chinese not plausible in the past. For example, the two news articles shown in Table 2 appeared separately in a Hong Kong and a Singapore newspaper on July 4, 1995. Both articles reported the same event of conflict between Egypt and Sudan following an assassination attempt on the life of Egyptian President Mubarak in Ethiopia. However, not only was the choice of words different, but the thrust and focus of the message also differed. These synchronous and homeothematic data can provide not only a good basis for comparative lexical studies in different Chinese speech communities, but also can enable researchers to conduct substantive stylistic and content analyses in social science research.

Table 2. Two pieces of synchronous and homeothematic news

Region	Hong Kong	Singapore
Date	4/7/95	4/7/95
Title (with mark-up)	<埃及px> <蘇丹px> <兩> <國> <關係> <惡化> @	<埃及px> <蘇丹px> <關係> <緊張>，<戰爭> <一觸即發> @
Original Text	埃及蘇丹兩國關係惡化@ 隨著與鄰國埃及的關係日益緊張，蘇丹政府威脅要切斷流往埃及的尼羅河水源，然而對於這個說法，一名國際水利專家卻認爲是天方夜談，絕無可能在不久將來實現。由於埃及政府指摘蘇丹日前策劃行刺其總統穆巴拉克，令到雙方關係急轉直下。	埃及蘇丹關系緊張，戰爭一觸即發@ 自埃及總統穆巴拉在埃塞俄比亞險遭暗殺事件發生后，埃及和蘇丹的關系陷入緊張狀態，戰爭可能一觸即發。昨日，蘇丹恫言如果埃及要用武力解決邊界糾紛，它將切斷流到埃及的尼羅河河水。較早時，埃及在報章警告："那些在蘇丹玩火的人……將逼我們走向對抗。"昨天數萬蘇丹人民軍在首都舉行大會，高舉AK-47沖鋒槍，詛咒穆巴拉，并誓言保衛國土。（路透社）
Content (with mark-up)	<隨著> <與> <鄰> <國> <埃及px> <的> <關係> <日益> <緊張>，<蘇丹政府h> <威脅> <要> <切斷> <流> <往> <埃及px> <的> <尼羅河h> <水源>，<然而> <對於> <這個> <說法>，<一> <名> <國際水利專家s> <卻> <認爲> <是> <天方夜談s>，<絕> <無> <可能> <在> <不久> <將來> <實現>。<由於> <埃及政府h> <指摘> <蘇丹px> <日前> <策劃> <行刺> <其> <總統s> <穆巴拉克pn>，<令到> <雙方> <關係> <急轉直下>。	<自> <埃及總統h> <穆巴拉pn> <在> <埃塞俄比亞px> <險> <遭> <暗殺> <事件> <發生> <後>，<埃及px> <和> <蘇丹px> <的> <關係> <陷入> <緊張> <狀態>， <戰爭> <可能> <一觸即發>。<昨日>，<蘇丹px> <恫言> <如果> <埃及px> <要> <用> <武力> <解決> <邊界> <糾紛>，<它> <將> <切斷> <流> <到> <埃及px> <的> <尼羅河h> <河水>。<較> <早> <時>， <埃及px> <在> <報章s> <警告>："<那些> <在> <蘇丹px> <玩火> <的> <人>……<將> <逼> <我們> <走> <向> <對抗>。" <昨天> <數> <萬> <蘇丹人民軍s> <在> <首都> <舉行> <大會>，<高舉> <AK-47 ix> <衝鋒槍s>，<詛咒> <穆巴拉pn>，<并> <誓言> <保衛> <國土>。（<路透社hx>）

5. LIVAC and Comparative Studies on Lexical Variation in Chinese

The material in LIVAC allows for a wide range of comparative studies on linguistic variation in Chinese. Some of the results of such comparative studies have been reported [T'sou 1993, 1995, 1996] and will be integrated with others in a future publication

[LIVAC 1997]. It is interesting to note other variations, such as the textual characteristics of newspapers in the different Chinese speech communities, in terms of the range of Chinese characters and words used, and differences in the nature of the language. For example, statistics concerning Chinese characters used in Hong Kong and Singapore, directly derived from the LIVAC corpus, offer an interesting comparison, especially in contrast to statistics concerning Taiwan extracted from the 1993 technical reports on the Academia Sinica corpus [CKIP 1993].

(a) According to the data available (see Table 3), the number of distinct Chinese characters used as a function of the percentage of corpus coverage is greatest in Taiwan, followed by Hong Kong, and then by Singapore. The gap between each consecutive pair is about 20%, and this number increases with the percentage of corpus coverage. These differences reflect the relative status and role of the Chinese language in the three communities: Taiwan is almost exclusively a mono-lingual Chinese speech community; Hong Kong is perhaps basically a mono-lingual society, though there is an unusual overlapping diglossic English-Chinese speech community as well [T'sou 1993]; and in Singapore, Chinese, being the language promoted by the government in a bilingual (but traditionally English dominant) setting, is used in a more basic form in news coverage [T'sou 1996]. It is interesting to note that the one thousand highest frequency characters in each location have more than 90% coverage of all the characters used in the media. While this may be true of characters, the same cannot be said of lexical items (see below and [T'sou 1995, 1996]).

Table 3. *Corpus coverage vs. the number of high frequency characters*

	No. of High Frequency Words		
Corpus Coverage	**Singapore**	**Hong Kong**	**Taiwan**
10%	9	10	12
20%	27	32	37
30%	54	64	71
40%	93	108	118
50%	144	168	179
60%	214	250	266
70%	318	373	395
80%	484	560	594
90%	792	911	983
100%	3613	4348	5666

(b) Some interesting trends can be observed on the range of Chinese words used in

different communities while noting that Chinese words might have resulted from different segmentation processes. Generally speaking, the segmentation rules adopted by LIVAC have resulted in finer granules of words. Thus, the 42,689 distinct words in LIVAC, for its entire Hong Kong corpus, is much smaller than that of the Academia Sinica 93 corpus, which reported 42,686 words having a frequency of occurrence of 3 or higher in its corpus. Therefore, it is interesting to note that, for a corpus coverage of 30% to 90%, the number of Chinese words used in Hong Kong exceeds that used in Taiwan, and that Taiwan uses a much wider repertoire of words.

Table 4. *Corpus coverage vs. the number of high frequency words*

	No. of High Frequency Words		
Corpus Coverage	**Singapore**	**Hong Kong**	**Taiwan**
10%	5	8	10
20%	24	33	35
30%	64	85	82
40%	144	188	178
50%	295	387	370
60%	566	774	725
70%	1075	1513	1331
80%	2113	3075	2447
90%	5043	7477	5005
100%	24967	42689	42686+

6. Conclusion

It is not surprising that the assumed uniformity of the Chinese language is open to question, given the size of the speech communities which use the language. The details on linguistic variations which have been uncovered are useful for a full understanding and appreciation of the Chinese language. Other trends are being explored following the successful implementation and application of segmentation efforts, and will be the subject of detailed study in the future. Additional development of automatic programs to identify homeothematic news stories in a synchronous time frame will make possible other studies based on content analysis, which could shed light on the rich cultural and social fabric to be found in individual Chinese speech communities.

Acknowledgments

The research reported here was supported by a grant from the Chiang Ching-Kuo Foundation for Scholarly Exchange in Taipei, and by a competitive Earmarked Research Grant from the Research Grants Council of the University Grants Committee of Hong Kong [(94)9040092]. The authors are grateful for the contributions made by W.M. Tham in Singapore, F.Y. Lin in Taiwan and W.F. Tsoi, E. Ma, C.S.Fan, G. Chan, A. Tang, C.A. Chin, S.K.Wong, L.Y. Wong, and K.L. Chan in Hong Kong.

References

Chen, C.Y., "Certain Lexical Features of Singapore Mandarin," in *New Papers on Chinese Language Use*, B. Hong (ed), Contemporary China Papers, No. 18, Australian University Press, 1984, pp. 93-104.

Chen, C.Y., S.F. Tseng, C.R. Huang and K.J. Chen, "Some Distribution Properties of Mandarin Chinese -- A Study Based on the Academia Sinica Corpus," in *Proc. 1st Pacific Asia Conf. on Formal and Computational Linguistics*, Taiwan, 1993, pp. 81-95.

Chen, P. P. S., "The Entity-Relationship Model: Towards a Unified View of Data," *ACM Transactions on Database Systems*, Vol. 1, No. 1, 1976, pp.9-36.

Sanders, G.L., *Data Modeling*, Boyd & Fraser Publishing Co., MA, 1995.

Report on Project LIVAC, Language Information Sciences Research Centre, City University of Hong Kong, 1997.

Tse, K.P.J., "Standardization of Chinese in Taiwan," *International Journal of the Sociology of Language*, No. 59, June 1986, pp. 25-32.

T'sou, B.K., "On the Linguistic Covariants of Cultural Assimilation." in *Anthropolical Linguistics*, Vol. 17, No. 9, 1975, pp. 445-465.

T'sou, B.K., "Distribution of Varieties of Chinese in the Pacific Region," in *Language Atlas of the Pacific Region*, compiled by S. Wurm and S. Hattori, Stuggart, 1983.

T'sou, B.K., "香港和中國大陸的一些語言現像 (Some Aspects of Language in Hong Kong and China)," in *Chinese Language Bulletin*, Chinese University of Hong Kong, No. 4, Sept. 1989, pp. 3-9.

T'sou, B.K., C.S. Lun, K.F. Liu, P.K. Wong and M. Sze, "漢語受事格位置初探 (A Preliminary Study on Object-Verb Construction in Chinese)," in *Proc. the World Conf. on Chinese Language Teaching*, Singapore, 1990, pp. 167-174.

T'sou, B.K., F. Liu, P.K. Wong, C.S. Lun and M. Sze, "香港電視新聞節目中的粵語與普通話用語初探 (Cantonese and Putonghau in Hong Kong TV News Programmes: A Preliminary Study in Language Use)," in *2nd Conf. on Cantonese and Other Yue Dialects*, Jinan Xuebao,

Journal of Jian Nan University, Guangzhou, Vol. 12, No. 1, 1990, pp. 70-76.

T'sou, B.K., "Some Issues on Law and Language in the Hong Kong Special Administrative Region (HKSAR) of China," in *Language, Law and Equality: Proceedings of the 3rd International Conference of the International Academy of Language Law (IALL)* held in South Africa, April 1992, (Ed. Karel Prinsloo, Yvo Peeters, Joseph Turi and Christo Van Rensburg), University of South Africa, Pretoria, 1993, pp. 314-331.

T'sou, B.K., "Lexical Innovation in Chinese: Some Methodological Consi-derations," paper given at *Symposium on Prisma Sprache: Chinesesche Versuche zur Bewaltingung Westlichen Gedankenguts*, Bad Homburg, July 1995.

T'sou, B.K., "星、港、台三地共時語料差異的分析與比較 (Lexical Variation in Singapore, Hong Kong and Taiwan)," Invited plenary session paper, *1st National Conf. on Language and Writing Applications*, National Commission on Language Reform, Dec. 1995.

T'sou, B.K., "Project LIVAC and the Exploration of Endocentric and Exocentric Developments in Lexical Variation in Different Chinese Speech Communities," Invited paper, *3rd Annual Conf. of the Y.R. Chao Centre for Chinese Linguistics*, University of California, Berkeley, March 1996.

T'sou, B.K, "香港大語料庫與星、台、滬用詞研究比較 (A Comparative Study of Lexical Usage in Singapore, Taiwan and Shanghai with the Hong Kong Language Corpus)," 第二屆語文現代化學術會議, 2nd Conference of Language Modernization, State Language Commission, 廣西師範大學, 桂林, Guilin, October 1996.

中文詞匯研究小組委員會編, "香港初中學生中文詞匯研究," 香港教育署出版, 1986.

北京語言學院語言教學研究所編, "現代漢語頻率詞典," 北京語言學院出版社, 1986.

劉源 等編, "現代漢語常用詞詞頻詞典," 宇航出版社, 1990.

台北中央研究院資訊科學研究所中文詞知識庫小組 (CKIP) 編, "新聞語料庫字頻統計表," "新聞語料庫詞頻統計表," "新聞常用動詞詞頻與分類," "新聞常用名詞詞頻與分類," 技術報告 93-01, 93-02, 93-03, 93-04, 1993.