Toward a Pan-Chinese Thesaurus

Benjamin K. Tsou and Oi Yee Kwong

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{rlbtsou, rlolivia}@cityu.edu.hk

Abstract

In this paper, we propose a corpus-based approach to the construction of a Pan-Chinese lexical resource, starting out with the aim to enrich existing Chinese thesauri in the Pan-Chinese context. The resulting thesaurus is thus expected to contain not only the core senses and usages of Chinese lexical items but also usages specific to individual Chinese speech communities. We introduce the rationale underlying the construction of the resource, outline the steps to be taken, and discuss some preliminary analyses. The work is backed up by a unique and large Chinese synchronous corpus containing textual data from various Chinese speech communities including Hong Kong, Beijing, Taipei and Singapore.

1. Introduction

Looking up a more recent edition of the Roget's Thesaurus (Kirkpatrick, 1987), one will find the word *subway* under head *624 Way*, among several other senses of the word. Under the same semicolon-separated group, one also finds *underground railway*, *tube* and *metro*. Similarly if one looks up WordNet[1] (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 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 *'di4tie3'* 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 *'jie2yun4'*. 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, no lexical resource in Chinese has yet achieved such comprehensiveness.

There are already English dictionaries and reference works based on, for example, the majority speech communities across the Atlantic and in the Pacific region including Australasia, Singapore and Malaysia, and India. However, nothing comparable exists for the Chinese language. In this work, we attempt to fill this gap by proposing a comprehensive Pan-Chinese lexical resource, starting with a Pan-Chinese thesaurus. The project takes advantage of 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*, such that it is not targeted at one particular community speaking the language and thus covering only language usage observed from that particular community. Instead, it 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 on *subway*, a lexical resource should be able to capture regional variation to be useful in a wide range of applications.

In Section 2, we will briefly review existing resources and related work. Then in Section 3, we will describe the design of our Pan-Chinese lexical resource, including its overall architecture, and the lexical relations and other linguistic features to be represented. In Section 4, we will further illustrate our idea of enriching existing thesauri in the Pan-Chinese context with an example. The proposed methodology will be discussed in Section 5, followed by a conclusion in Section 6.

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).

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 fortunately and 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.

[1] http://wordnet.princeton.edu/

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 arisen as a result of the individualistic evolution and development trends 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. It was first published in the 80s and based exclusively on Chinese as used in post-1949 Mainland, and thus for the *subway* example above, the closest word group found is 火車 *'huo3che1'*, 列車 *'lie4che1'* (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 aspects of the core common language in constructing a lexical resource; and to incorporate useful information relating to location-sensitive linguistic variations. In this way, divergent and convergent trends could be represented within a Pan-Chinese context.

3. Design

The proposed Pan-Chinese lexicon is expected to capture not only the core senses of lexical items but also senses and uses specific to individual Chinese speech communities. Our work is thus backed up by a large (about 90 million characters) and unique synchronous Chinese corpus, which will be introduced in the following subsection. In this section, we will also outline the basic architecture of the proposed resource and the lexical relations to be represented in the thesaurus, as well as discuss the nature of regional variation of lexical items.

3.1. The LIVAC Synchronous Corpus

LIVAC[2], which stands for Linguistic Variation in Chinese Speech Communities, 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 contains newspaper articles collected synchronously and regularly from six Chinese speech communities, including Hong Kong (HK), Beijing (BJ), Taipei (TP), Singapore (SG), Shanghai and Macau. Texts collected cover local news, international news, sports news, entertainment news, and financial news. The texts are segmented into words which are in turn tagged with part-of-speech categories.

The corpus thus provides a unique and authentic textual database for lexical acquisition. Qualitative and quantitative analyses with the corpus data on the sociolinguistic aspects of Chinese language in the various communities, as well as lexical comparisons, have revealed insightful and interesting linguistic patterns and differences. For instance, the literal and physical sense of 打造 *'da3zao4'* (to fabricate) was first found to have shifted to figurative use by taking on non-concrete objects in the Taipei data.

3.2. General Architecture

The 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.

As a first step, we are working toward a Pan-Chinese thesaurus by acquiring near-synonyms from LIVAC, integrating them with and thus enriching existing thesauri. We plan to adopt a network structure for our thesaurus data which is in a way similar to WordNet (Miller et al., 1990). A network structure is preferable because it is intuitively sound for the representation of semantic relations (cf Quillian's (1968) semantic memory) and will be easily usable for semantic analysis and inferencing. 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). Figure 1 shows the schematic organization of the Pan-Chinese semantic lexicon.

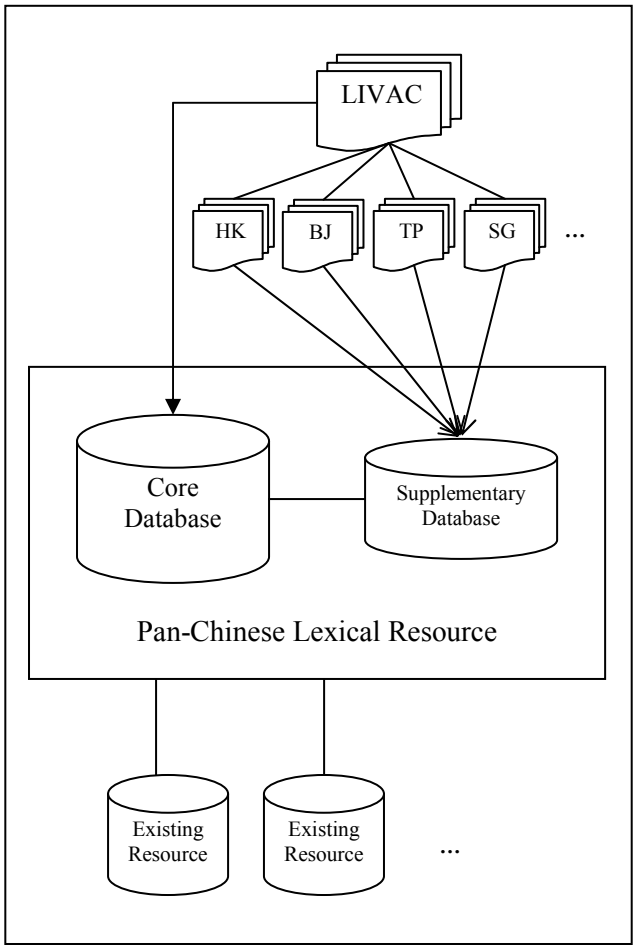

Figure 1: Design of the Pan-Chinese Lexical Resource

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

3.3. Lexical Relations and Regional Variation

In the thesaurus, as in other resources of a similar kind like Roget's Thesaurus and WordNet for English, and Tongyici Cilin for Chinese, lexical items will be organised into a hierarchical network based on conceptual relatedness. In addition, we will map out the semantic relations of *synonymy*, *hypernymy*, *hyponymy* and *antonymy*, with word samples drawn from the above-mentioned authoritative 90-million-character LIVAC synchronous Chinese corpus. In particular, in the fuller resource, we will attempt to classify each set of related lexical items or near-synonyms into one or more of the following categories to indicate the source of differentiation among members of the set:

(i) *social registers* – words might have similar meanings but might often be used in association with particular registers, e.g. 座駕 *'zuo4jia4'* / 汽車 *'qi4che1'* / 車子 *'che1zi5'* all mean "car", but apparently the first is more formal than the second, which is in turn more formal than the third;

(ii) *adaptational differences* – for concepts of foreign origins, their lexicalisation in Chinese might be different across various communities in terms of phonetic and/or semantic adaptation, e.g. "taxi" is called 的士 *'di2shi4'*, 德士 *'de2shi4'*, 計程車 *'ji4cheng2che1'* and 出租車 *'chu1zu1che1'* in Hong Kong, Singapore, Taipei, and Beijing respectively;

(iii) *dialects* – near-synonyms might arise as a result of dialectal difference, e.g. 收工 *'shou1gong1'* and 放工 *'fang4gong1'* are Cantonese terms for "to get off work" and are therefore often only found in Hong Kong Chinese, whereas the synonymous term in Modern Standard Chinese is 下班 *'xia4ban1'* which therefore is the normal form found in Mandarin-speaking communities; and

(iv) *cultures* – lexical difference across Chinese speech communities might also be a result of the historical and therefore cultural influence of individual communities, e.g. the "police force" is known as 警察 *'jing3cha2'*, 公安 *'gong1an1'* and 司警 *'si1jing3'* in Hong Kong, Beijing and Macau respectively.

Regional variation does not only show in different lexicalisation, but also surface in different senses of the same word forms. For instance, the verb 拉 *'la1'* normally refers to the physical action of pulling in Modern Standard Chinese, but is also used in Hong Kong data to mean "arrest", which in turn is mostly rendered as 拘捕 *'ju1bu3'* in other places and as a more formal way of putting it.

Even near-synonyms might have different behaviours in various communities. The semantic shift of 打造 mentioned in Section 3.1 is an example. For comprehensiveness and usefulness, the proposed Pan-Chinese lexical resource will also capture other linguistic features including the relative distribution as well as collocation patterns, as exhibited in the LIVAC data.

4. An Example

In this section, we present an example to further illustrate the idea of the Pan-Chinese thesaurus, with a view to enrich existing thesauri. The word 開 *'kai1'*, for instance, is a very versatile word. It appears under the following 13 semantic heads in Tongyici Cilin[3]:

(i) Fa10 挖 剔 鉗
… 扒(~土), 打(~井), 開(~河;~井) …

(ii) Fa31 開 關
… 開, 打開, 張開, 啓 …

(iii) Hb06 射擊 空襲 爆破 刺殺
… 發射, 開(~槍), 發(~炮) …

(iv) Hc06 建立 設立 創立 命名
… 開設, 開辦(~訓練班), 設(~攤), 開(~店) …

(v) Hc22 處罰 檢討
… 革出(~教會), 革除, 開(資本家隨便~掉工人) …

(vi) He10 徵收 交納 支付
… 支(~些錢給他), 出(量入爲~), 開(~工錢) …

(vii) Hf01 駕駛 駕御
… 駕駛, 駕(~飛機), 開(~摩托車) …

(viii) Hg11 揮筆 記錄 留言 附筆
… 題(~字), 修(~家書一封), 開(~發票) …

(ix) Ia10 沸騰 蒸發 溶解 融化
… 開(水~了), 沸(油~了), 滾(粥~了) …

(x) Ia11 冷卻 結冰 解凍
… 化凍, 開河, 開化, 開(等河~了坐船走) …

(xi) Ib21 開花 結果 凋謝
… 開(百花盛~), 放(百花齊~) …

(xii) Ie13 實行 舉行
… 舉行, 召開, 開(~運動會), 做(~生日) …

(xiii) Ig01 開始 結束
… 始(不自今日~), 開(~演), 啓(~行;~用) …
… 起行, 出發 … 開拔, 開(部隊已經~走了) …

Apparently most core senses of 開 are covered above. However, some other senses commonly found in our corpus are missing, e.g. distant: (站)開 *'zhan4kai1'* (to stand further away); portion (of a newspaper): (四)開 *'si4kai1'* (a paper symmetrically divided into 4 portions). In addition, some region-specific uses of the word are not found. For example, the Mandarin usage of 開(三個饅頭) *'kai1 (san3ge5man2tou5)'* (to eat (3 buns)) is not found, and should probably be grouped under *head Fc06* 吃 嚼 咽 吮 喝. Similarly, the Cantonese usage of 開(牛奶) *'kai1 (niu2nai3)'* (to make (a glass of milk)) should ideally also come under the stirring action in *head Fa26* 攪拌 摻和 拺.

Meanwhile, the usages in (v), (ix) and (x), for example, are apparently more prevalent for Mandarin speakers. Other alternatives such as 撤職 *'che4zhi2'* or even 炒 *'chao3'* for dismiss might contribute to the usefulness of the thesaurus.

5. Methodology

We start with preliminary analysis on the most frequently used 50K Chinese lexical items in the corpus and focus on the approximately 35K content words therein, which would be of greater interest in terms of their semantic contents than function words and proper nouns. Analysis of related lexical items will be in terms of (i) the morphological structure of the words, including whether a word is a phonetic adaptation of a foreign concept, (ii) the word collocation patterns, (iii) the syntactic contexts

[3] The 13 senses in English are: (i) to dig (a well), (ii) to open (a container), (iii) to fire (a pistol), (iv) to establish (a company), (v) to dismiss (a worker), (vi) to release (salary), (vii) to drive (a car), (viii) to issue (a cheque), (ix) (water) boils, (x) (a river) defrosts, (xi) to blossom, (xii) to hold (a meeting), and (xiii) to begin (an event).

embedding the words, (iv) the semantic contexts, and (v) semantic shift. Some results of the preliminary analysis are presented in Section 6.

Based on the analysis we will look for useful features for the automatic extraction of more related lexical items from the corpus, and then experiment with mature computational techniques (e.g. Caraballo, 1999; Riloff and Shepherd, 1999) to do the extraction in a larger scale.

We will also examine how our results can be integrated with existing resources like the Tongyici Cilin to provide a more comprehensive lexical resource.

6. Preliminary Analysis

We started with in-depth linguistic analyses for several seed sets of related lexical items sampled from the LIVAC corpus (Cheng et al., 2004; Kwong and Tsou, 2005). The analysis of a set of related words pertaining judgement (裁定 *'cai2ding4'* / 裁決 *'cai2jue2'* / 判決 *'pan4jue2'*) in the Pan-Chinese context reveals that there are far more nominal usages than verbal usages of 裁定 in Beijing than in any other communities. Moreover, Singapore data was found to rely on 裁定 to cover the conclusion (verdict) and the consequence (sentence and order) simultaneously, while words like 判刑 *'pan4xing2'*, 判監 *'pan4jian1'*, 判罰 *'pan4fa2'*, 判囚 *'pan4qiu2'* and 判處 *'pan4chu3'* (all related to sentencing) are relatively more abundant in Hong Kong or Taipei data than in Singapore data, suggesting that Hong Kong and Taipei tend to distinguish between the verdict and the sentence more clearly.

Another analysis on a group of reportage verbs, including 說 *'shuo1'* (to say) / 表示 *'biao3shi4'* (to express) / 指出 *'zhi3chu1'* (to point out) / 稱 *'cheng1'* (to claim), shows that while they are frequent in all Beijing, Taipei, and Hong Kong, the relative distribution could be very different and the same verb can have very distinctive usages across different regions. The subtle difference lies in their subcategorisation frames, thematic roles they take, selectional restriction, and the like. For example, pronominal and inanimate agents often collocate with 說 and 稱 but seldom with 表示. On the other hand, 說 and 稱 differ in their syntactic realisation of semantic roles. The latter seldom finds itself collocating with pronominal subjects for agent and direct quotations for patient.

Thus while most automatic lexical acquisition methods rely on the distributional similarities of closely related words, it also remains for us to further investigate how we might extract those which differ considerably in their syntactic realisation.

7. Conclusion

In this paper, we have proposed a Pan-Chinese lexical resource which captures not only core word senses and usages of Chinese, but also region-specific "marginal" and "sublanguage" uses. The project is based on a unique and large synchronous corpus of Chinese. Work is underway toward a Pan-Chinese thesaurus, an essential part of the lexical resource, aiming at enriching existing Chinese thesauri which are usually limited in regional variation.

8. 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 thank the anonymous reviewers for comments.

9. 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.

Cheng, C.M., Kwong, O.Y. and Tsou, B.K. (2004) Pan-Chinese Variation on Verbal Synonymy: A Study of Common Reportage Verbs in News Texts. In *Proceedings of the 5th Chinese Lexical Semantics Workshop*, Singapore, pp213-219.

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

Kwong, O.Y. and Tsou, B.K. (2005) A Synchronous Corpus-based Study on the Usage and Perception of Judgement Terms in the Pan-Chinese Context. *International Journal of Computational Linguistics and Chinese Language Processing, 10(4)*: 519-532.

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 / 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 on-line lexical database. *International Journal of Lexicography, 3(4)*:235-244.

Quillian, M.R. (1968) Semantic memory. In M. Minsky (Ed.), *Semantic Information Processing*. Cambridge, MA: MIT Press.

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 Lai, T.B.Y. 鄒嘉彥、黎邦洋 (2003) 漢語共時語料庫與信息開發. In Xu et al. 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.