Quantitative Criteria for Computational Chinese Lexicography

A Study Based on a Standard Reference Lexicon for Chinese NLP

Topic Areas: (d) electronic dictionaries, (h) large corpora

Chu-Ren Huang, Zhoa-ming Gao, Claude C.C. Shen, and Keh-jian Chen

Academia Sinica

Email:hschuren@ccvax.sinica.edu.tw

Fax:(02)2788-1638

Abstract

The construction of a standard reference lexicon for Chinese NLP involves two fundamental issues in computational linguistics: the definition of a word and the principled delimitation of the lexicon. We argued that such reference lexicons must be judged by their cross-domain portability, expressive adequacy, and reusability. Thus principles for lexical selection must also be driven these criteria.

This paper reports the approach and result of our construction of a standard reference lexicon for Chinese NLP, which also serves as the empirical basis for a segmentation standard. Our approach uses a mixture if stochastic and heuristic steps. First, a reference corpus is selected and lexical entries are automatically extracted from it based on statistically significant threshold. Second, the coverage of the automatically extracted lexicon is enhanced by conceptual primes as well as by comparative studies of MRD's from different Chinese speaking communities. We show the satisfactory coverage of the resultant lexicon by testing it with randomly accessed texts from the web.

1 Introduction

Since words are not conventionally marked in Chinese texts, segmentation is a pre-requisite step for Chinese NLP and setting a standard to define and measure segmentation results becomes necessary both for evaluation and for resource-sharing. However, as noted by standard-setters both in Mainland China (Liu et al. 1994) and in Taiwan (Huang et al. 1996), no segmentation standard can be successfully implemented and evaluated until it is accompanied by a wide-coverage reference lexicon. Huang et al. (1996) argued that segmentation standards must include a sharable and adaptable lexicon in order to apply across variations such as domain, genre, and time. On the other hand, for data-sharing, a standard lexicon is essential to ensure that texts from different sources can be uniformly tokenized. Thus, there is

a consensus that an empirically compiled reference lexicon is an indispensable part of a Chinese segmentation standard (e.g. Lin and Miao 1997, Sun and Zhang 1997). An additional benefit of such a lexicon is that it can be shared without much additional cost and thus saves NLP researchers the time and cost in building this essential infrastructure. To serve the above dual purposes, this standard lexicon must be selected in a principled way in order to best test validity and usefulness.

However, even though computational lexicography offers a rich literature on the structure and content of a lexical entry, there is hardly any discussion of a principled way of lexical entry selection (Armstrong-Warwick 1995). We only see discussion in the content of a terminology lexicon (Nagao 1994) or a reference segmentation lexicon (Liu et al. 1992). Both assume that the lexicon is built from scratch. We suggest that there are three criteria to judge the merit of a lexicon: reusability, expressive adequacy, and domain-portability. Based on these three criteria, we propose a principled way to construct a standard reference lexicon for Chinese NLP.

2 Criteria for Selection of Lexical Entries

2.1 Word, Segmentation Unit and Lexical Entry

Determining whether a string is a (new) word is trivial in many languages such as English. However, it is not easy in Chinese because of the lack of conventional demarcation and of native speakers' consensus of what is a word. Once a string is identified as a unit, a further decision needs to be made as to whether it should be listed in the lexicon (e.g. Wang et al. 1994).

In this study, we stipulate that all entries must be segmentation units defined in

Huang et al. (1997a & b). Notice that even though Huang et al. Propose to take the notion of linguistic word as the theoretical foundation of the definition of segmentation unit, it is obvious that certain non-words, such as (derivational) affixes, must also be treated as segmentation units. Thus they must be listed in the reference lexicon for segmentation.

The motivation of such a stipulate is two fold. First, it ensures uniformity of the segmentation criteria and the reference database within the segmentation standard. Second, this allows the reference lexicon to list non-words such as derivational affixes, and thus will provide crucial information to account for the productive morpho-lexical processes.

2.2 Reusability: Corpus Base of Lexical Selection

That the corpus is the best source of lexical entries has been the cornerstone of recent developments of corpus linguistics (e.g. Sinclair 1987). Making a balanced corpus as the basis of a standard reference lexicon also makes it possible to automatically update the lexicon for different domain or for language changes. Either a monitor corpus will be maintained to indicate any change in the language, or comparable corpus from separate domains can be maintained, any new entries can be extracted by the same automatic procedure to augment and revise the standard set. Our current lexicon is based on the Sinica Corpus (Chen et al. 1996), a tagged balanced corpus of Taiwan Mandarin Chinese containing 5 million words. 146,876 different words appear in the corpus. The number of lexical entries defined by frequency threshold of 1 to 10 are as follows:

(1) Number of Lexical Entries as Defined by Sinica Corpus Frequency

frequency threshold	number of word types
1	146,876
2	*84,309*
3	*63,421*
4	*52,571*
5	*45,443*
6	*40,395*
7	*36,392*
8	*33,301*
9	*30,701*
10	*28,564*

Our data shows that the frequency thresholds of 10 and 4 correspond to sharp increase in the number of lexical entries. Thus these are the two thresholds that will be adopted later in this work.

2.3 Expressive Adequacy: Conceptual Primes and Lexical Selection

Selecting lexical entries by frequency threshold based on corpus calculation is a dependable way to ensure relatively high coverage of the lexicon. However, since lexical information is nit available in NLP unless it is encoded in the lexicon, high coverage does not necessarily translate into successful application if conceptually crucial items are missing. Thus, we propose that a standard reference lexicon must achieve expressive adequacy. Our hypothesis is that such adequacy can be ensured when entries representing conceptual primes are exhaustively included.

The conceptual primes that we adopt are the 3,922 covering terms of Tongyici Cilin (Mei et al. 1983, CILIN hereafter), the most widely used thesaurus in Chinese NLP. We treat them as if they are covering terms in semantic fields, assuming these terms alone will be adequate to express concepts represented by embedded terms in their fields. Thus a lexicon containing all these terms will be expressively adequate.

A possible objective to adopting such an heuristic method independent of corpus – based stochastic approaches is that the same goal could be achieved without the heuristic. In the other words, is there any evidence to prove that these conceptual primes cannot satisfactorily extracted from corpora.

Diagram 1 shows the frequency/rank correlation of the CILIN conceptual primes based on their occurrences in Sinica Corpus. If conceptual primes were to be reliably extracted from corpora, they must fall (almost) exclusively in mid to high frequency rank. However, diagram 1 follows Zipf's law. In the other words, these conceptual primes areas widely distributed as other lexemes. Any corpus-based frequency threshold will unfortunately exclude the lower frequency conceptual primes.

One possible explanation for the Zipf's Law like distribution of the semantic primes is that a complete conceptual system needs to express all concepts regardless of their frequencies. The less frequently used semantic primes are those involving surprises or rarity, both are fundamental concepts. In the other words, a complete set of semantic primes necessarily contain less frequent words and their distribution should reflect the distribution should reflect the distribution of all meaning expressible by the language (i.e. the lexicon).

In fact, only 3,501 of the CILIN covering terms occur in Sinica Corpus, meaning that 421 terms are missing. These missing terms cannot be attributed solely to the lexical difference between Mainland China and Taiwan. Two authoritative dictionaries that consulted corpus extensively also do not enter all the CILIN (primes. The 57,624 entry Xiandaihanyu Cidian (XHCD hereafter) lacks 241 of them while the 39,025 entry Segmentation Standard Lexical List (Liu et al. 1994, GB hereafter) misses 546 of them. There does not seem to be a correlation between the degree of human intervention with the completeness of conceptual primes though. XHCD is

compiled by linguists who consulted corpora, while GB is extracted from a corpus and augmented with thought-up lexical items.

In diagram 2, an addition test is conducted on the distribution of these conceptual primes. The diagram shows the number of conceptual primes every 1,000 words in Sinica Corpus ordered according to frequency rank. As suspected, a high proportion of the most frequent words are conceptual primes (382 of the first 1,000), while the proportion descends dramatically. The diagram shows the slope smoothes at around rank 1,500 and levels well before rank 10,000.

Two important pieces of information can be inferred from diagram 2. First, it offers an intuitive support of the reliability of the CILIN primes. Since conceptual primes are the most economic (and often necessary) way to express ideas, they are more likely to be frequently used. Thus, we expect a valid set conceptual primes to be dominated by high frequency words. The CILIN distribution confirms such prediction.

Second, the steep descend and quick leveling suggests that it will be impractical to discover conceptual primes with pure stochastic approach. Since these conceptually primary terms are sparsely distributed in mid to lower frequency range, it would be quite impossible to achieve any reasonable recall and precision at the same time. In other words, for the moment at least, conceptual primes must be acquired independent of a corpus.

2.4 Portability: Bootstrapping with Existing Lexicons

It is impossible for a corpus, with finite total words, to cover all possible topics, genre etc. Hence it is most likely that some significant lexemes are not represented in a corpus. In other words, how can a standard lexicon be portable among all domains given the fact the corpus it based on does not contain texts from all possible domain?

This problem could be aggravated if a corpus is relatively small and geographically restricted.

The case is even worse for a Chinese lexicon because of the fact that there exist substantial lexical differences between Mainland and Taiwan Mandarin. Thus it would be futile to construct a corpus that could represent both dialects. However, it is also well-known that mutual lexical borrowings are easy and frequent contacts. Thus any purely Taiwan or Mainland corpus faces the dilemma of under-representing a critical segment of lexemes.

To solve this dilemma, we propose to bootstrap with lexicons. We consult the entries of five lexicons, including two each from PRC and Taiwan, as well as one from the U.S. The two Mainland lexicons: List of Frequently Used Modern Mandarin Words for information Processing (Appeared in Liu et al. 1994, referred to as **GB** hereafter), and Xiandai Hanyu Cidian (Chinese Academy of Social Sciences 1996, referred to as **XH** hereafter). The two Taiwan lexicons are: the Chinese Knowledge Information Processing Electronic Lexicon of Academia Sinica (last updated 1996, referred to as **CKIP** hereafter), and the on-line version Revised Revision of Mandarin Chinese Dictionary by the Council on Mandarin Chinese of the Ministry of Education (1997 version, referred to as **RMCD** hereafter). Lastly, the **ABC** Chinese-English Dictionary (DeFrancis 1997, referred to as ABC hereafter) not only represents the perspective of a language learner but also offers a perspective not dictated solely by linguistics experience in one single area.

(2) Number of Lexical Entries in the Five Dictionaries

Dictionaries (year of compilation)	**Number of Entries**
CKIP 1996	78,323
RMCD 1997	156,710
XH 1996	56,162
GB 1993	39,459
ABC 1997	70,325

Our claim is that comparing entries from compiled lexicons allows us to tap existing knowledge and labor-intensive resources. The decision to include a lexical entry in a lexicon reflects the collective knowledge of (at least a good number of) native speakers and is at least as valuable as un-processed raw corpus data.

2.4.1 Towards a Formal Definition of Lexicon Similarity

In this section, we will try to set a principled way of comparison of lexicon as well as to interpret the important of repeated occurrence of an entry in different lexicons. As shown by (2), the sizes of the five lexicons vary greatly, from just under 40 thousand entries to over 156 thousand entries. Since these lexicons are describing the same language, they should in principle have very similar entries. Thus the two questions that one must ask ate 1) roughly speaking, are the smaller lexicons subsets of the larger lexicons? 2) are the lexicons compiled in the same geographic area more similar to each other? To answer the two questions, we start by finding out the coverage rate of each dictionary pairs. The coverage of dictionary A over dictionary B is defined as

(3) Coverage of Dictionary A over Dictionary *B*

$Cov_{A/B} \overset{def}{=}$ Number of entries in the intersection of A and B/Total number of entries in B

Based on the above definition, the coverage rate among the five dictionaries are calculated as below:

(4) Coverage Among the Five Dictionaries

B \ A	CKIP	RMCD	XH	GB	ABC
CKIP	100%	68.89%	45.85%	35.94%	50.34%
RMCD	34.42%	100%	29.42%	19.72%	30.76%

XH	63.94%	82.10%	100%	50.92%	75.36%
GB	71.33%	78.30%	72.48%	100%	79.44%
ABC	56.07%	68.58%	60.18%	44.58%	100%

Take note that the above definition of coverage is dependent on the size of the lexicon. That is, mathematics speaking, a similar lexicon cannot have a good coverage of a bigger lexicon since it cannot cover of a larger lexicon over a smaller one is not especially high. For instance, although **RMCD** is almost four times as big as **GB**, it only covers 78.30% of the later. This and the wide range coverage numbers suggests that we need a better criterion for dictionary similarity. We cannot ignore the fact that number of entries is a very important feature of any lexicon. However, to make sure that extreme difference in sizes do not skew the similarity between lexicons, we propose that mutual coverage as a good measure of lexicon similarity.

(5) Mutual Coverage of Two Lexicons A and B

$$\mathbf{Mcov_{A,B}} \overset{def}{=} \mathbf{Cov_{A/B} + Cov_{B/A}/2}$$

Based on the above definition, mutual coverage among the five lexicons are given below from the highest mutual coverage rate to the lowest.

(6) Mutual Coverage among Five Lexicons (descending order)

$\mathbf{MCov_{ABC,XH}}$	67.77%
$\mathbf{MCov_{ABC,GB}}$	62.07%
$\mathbf{MCov_{XH,GB}}$	61.70%
$\mathbf{MCov_{RMCD,XH}}$	55.76%
$\mathbf{MCov_{CKIP,XH}}$	54.90%
$\mathbf{MCov_{CKIP,GB}}$	53.64%
$\mathbf{MCov_{CKIP,ABC}}$	53.21%

$\mathbf{MCov}_{\mathbf{RMCD,CKIP}}$ 61.65%

$\mathbf{MCov}_{\mathbf{RMCD,ABC}}$ 49.67%

$\mathbf{MCov}_{\mathbf{RMCD,GB}}$ 49.01%

The above result confirmed our suspicion that **ABC**, **XH**, and **GB** are more similar to one another. This is because these follow the PRC usages predominantly, including ABC, although it is compiled in the States. However, **RMCD** and **CKIP** do not show the same degree of similarity. As a matter of fact, all the other three lexicons are more similar to **CKIP** than **RMCD** according to this measure. In other words, it is more than simply geo-political influence that determines the similarity of the lexicons. The criteria of lexical selection as well as the topic areas covered will play a crucial role too. **RMCD** has a selection criterion that is quite different from the other lexicons, that is it tries to be exhaustive without being sensitive whether an entry is commonly used by the speaking community. This may contribute to the reason why it appears to be the most different from the other four lexicons in our calculation. Another way to check the similarities of these five lexicons is to find out how many entries are shared by them. We found that all together there are 206,802 different word types (i.e. entries) recorded, and among them only 21,655 entries are entered in all five lexicons.

(7) Number of shared entries

a. shared by all five lexicons 21,655

b. shared by (at least) four lexicons 35,924

c. shared by (at least) three lexicons 54,111

d. shared by (at least) two lexicons 82,332

We believe the above data points to a definition of a standard core lexicon that is used most by most Chinese in most contexts. As we see, any two lexicons are only 50% to 60% similar. We further see that the number of entries that all lexicon

compilers agree upon is only 21,655. This is only a small fraction of all number of entries in each lexicon.

2.4.2 Why Lexicons Differ: the emergence of a core lexicon

The above study of different lexicons as well as earlier computational lexicography studies based on corpus suggest that lexicographers as well as corpora are biased. That the core lexicon entries tend to be covered by different corpora and different lexicographers. But there will be a lot of disagreement among corpora as well as lexicographers when more peripheral entries are being chosen. Thus, we can see a core lexicon emerging when we compare different authoritative lexicons as well as consult reliable large corpus. In the next section, we will propose a principled way to construct standard lexicons based both on dictionary and corpus knowledge so that the bias of each methodology can be canceled and valuable information from each approach can be utilized.

3 Principle and Methodology Towards a Standard Reference Lexicon

To meet the criteria of reusability, expressive adequacy, and cross-domain portability, we combine a three step algorithm for constructing a standard reference lexicon for Chinese NLP. First, lexical entries are automatically extracted from a balanced tagged corpus if their frequencies are higher than a stochastically determined threshold. The corpus-based generation allows automatic updating and adaptation to specific domains. Second, the automatically generated lexicon is augmented with a small set of conceptual primes to ensure expressive adequacy. Last, it is further augmented with entries obtained form intersection of 5 lexicons form different

sources to ensure cross-domain portability.

First, we define three levels of standard lexicons. The **Core Lexicon** is the most stable part of the language. It will be used regardless of geographic area, topic, media, style, genre, etc. In other words, it is the core of the segmentation standard that will be portable through different uses and through a reasonable duration of time. Second, the **General Lexicon** is a superset of the core lexicon. The extension over the core lexicon allows it to give better comprehensive coverage of text in general domains (such as newspapers or general textbooks). Last, the **Reference Lexicon** is an open set that is also the superset of the general lexicon. We want to included all lexical entries that are arrested words currently being used in the language (and are also segmentation units) to be listed in the reference lexicon. Ideally, the reference lexicon will have attribute attached so that special sub-lexicons can be automatically extracted for the special uses. But such annotation and expansion of the reference lexicon will involve voluntary cooperation of users from all different backgrounds. Right now, we envision the reference lexicon as an open set maintained virtually by R.O.C. Computational Linguistics Society. Any new lexical items nit covered by the current version of the reference lexicon will be reported on-lined. A team of experts will double-check that the reported new entries meet the required criterion of being a segmentation unit, and admit the entry to the reference lexicon.

On the other, the core and general lexicons will be maintained and updated periodically, perhaps every 3 to 5 years. The update will be based on corpus data as well as revisions on the dictionaries consulted. The update will allow the two lexicons to keep with linguistics changes, which is most evident in the area of lexicon.

3.1 Extraction of the Standard Lexicon: a hybrid approach

Our current algorithm for extracting the three levels of standard lexicons are:

(8) Core Lexicon

Entries must be listed in all five lexicons (**ABC**, **CKIP**, **GB**, **RMCD**, and **XH**), as well as occur for at least 10 times in the Sinica Corpus.

(9) General Lexicon

Entries must be listed in at least three of the five lexicons (**ABC**, **CKIP**, **GB**, **RMCD**, and **XH**), as well as occur for at least 4 times in the Sinica Corpus.

(10) Reference Lexicon: Entries must either

a. be listed in at least three of the five lexicons (**ABC**, **CKIP**, **GB**, **RMCD**, and **XH**); or

b. be listed in at least one of the five lexicons (**ABC**, **CKIP**, **GB**, **RMCD**, and **XH**) and occur at least once in the Sinica Corpus; or

c. be listed as one of the semantic primes in *Tongyici Cilin*.

Please note that the heuristic for the reference lexicon above attempts to extract the largest list possible of legitimate entries without human intervention. The three disjunction conditions are three different ways to make sure that an entries is indeed a lexical entry and segmentation unit in the language and not just a careless mistake of a lexicographer or an accidental error in a corpus. As mentioned above, it will then require continuing human intervention in the future to maintain the growth of the reference lexicon. The number of entries thus collected are listed in (11).

(11) Number of Entries of

a. **Core Lexicon: 13,049**[1]

b. **General Lexicon: 26,443**

c. **Reference Lexicon: 81,787**

4 Verification and Expendability

To verify that our standard reference does not meet the requirements set out by the three criteria, we will do both internal and external tests. Tests are performed with an automatic segmentation procedure to determine coverage of the lexicon of all words appearing in their language. Internal tests will be performed in texts extracted from Sinica Corpus, which are marked with topic, genre, style, media etc. Our aim will be to ensure that consistently high coverage is achieved across all possible variations. External tests will be performed with texts not included in Sinica Corpus, especially texts from Mainland China as well as texts extracted from WWW.

4.1 Verification of the Versatility of the Core Lexicon

We have mentioned above that the most important attributed of the core lexicon is its versatility, i.e. that it will be least sensitive to the change of texts and will still offer the same coverage. To test this requirement, we use all the texts in Sinica Corpus to as internal test set. As described in Chen et al. (1996), the over 500 texts in the Sinica Corpus are given textual mark-up in five different dimensions: Spoken/Written, Topic, Media, Genre, and Style. In each dimension, there are further divisions. For instance, Topic attributed included: Philosophy, Psychology, Chemistry, Society Culture, International Relationship etc. And Media attributed included Newspapers, Academic Journals, Audio-Visual etc. Thus we will be able to check the coverage of the core lexicon with regard to the dimensions of variations. The baseline lexicon

[1] The number of Core Lexicon is comparable to the theory of "詞涯七千" (Cheng, 1998).

we use to compare in this case is the 13,049 most frequent words in the Sinica corpus. In other words that are known to have the highest coverage of the collective texts. Thus the fact that the core lexicons has more stable coverage than this set of words will be one of the strongest possible evidence to show the versatility of the core lexicon. First, we adapt the definition of coverage given in (3) define the lexical coverage of a text.

(12) Lexical Coverage of a Text by a Lexicon L

$\text{LexCov}_L \overset{def}{=}$ Number of L's entries that appear in the text/Total number of word types in the text

Sine the baseline set contains the most frequent words of the corpus, it is mathematically impossible for the core lexicon to have higher coverage. So what we need to show crucially is that core lexicon will have a more stable coverage regardless of the nature of texts, given that its coverage is not too much lower than the most frequent word list. The statistical method we choose is the standard deviation of the coverage among all texts.

(13) Core Lexicon vs. Most Frequent Words

		Lexical Coverage	Standard Deviation
a) Spoken	Core	62.728%	6.05422%
	HiFre	76.7088%	6.69072%
b) Written	Core	57.434%	6.63843%
	HiFre	69.1228%	8.40772%
c) Topic	Core	53.1445%	2.47149%
	HiFre	64.5383%	2.71369%
d) Media	Core	58.1081%	3.39812%

	HiFre	69.6637%	3.93821%
e) Genre	Core	58.2538%	3.26635%
	HiFre	69.2185%	3.68371%
f) Style	Core	56.8369%	0.958196%
	HiFre	68.0522%	0.796978%

Take note that the above average is calculated based on the parameters within each dimension. For instance, the average and standard deviation under Topic is calculated based on the average coverage of the 56 topic divisions. The coverage of each topic division is in term calculated based on the coverage of all the texts assigned to that topic division. Thus what the test shows us is the performance of the core lexicon when confronted with variations in 5 different dimensions. The result is very reassuring in that although the lexical coverage of the core lexicon is slightly lower than the most frequent word list, as expected; its standard deviation is almost always lower than that of the most frequent words. And in the four dimensions where the core lexicon has a lower standard deviation, the difference is statically significant. The only case where the most frequent words have a lower standard deviation is in the Style dimension. However, in this case both standard deviation are very low and the difference even lower (only about 0.16%). This actually suggest that lexical coverage does not differ when the style (e.g. descriptive vs. expository etc.) changes.

In addition to internal tests on texts in the Sinica Corpus, we also did external tests with texts extracted from WWW. Since the Sinica Corpus is based in Taiwan, we tried to extract texts from the PRC. One caution with the external test is that the texts are automatically segmented, and were not manually checked like the Sinica Corpus. Thus the segmentation result may not only be 90%-95% correct. The test size is about 100,000 words.

(14) Lexical Coverage: external test

	Lexical Coverage	Standard Deviation
Core	59.1798%	3.68189%
HiFre	64.7171%	4.34954%

As expected, the standard deviation of lexical coverage by the core lexical is still significantly lower than that of the most frequent words from the Sinica Corpus. What is also reassuring is that the lexical coverage remain reliable at around 60% for the external texts. Since these are more frequent words, the textual coverage (i.e. coverage of tokens) is actually around 80%.

To sum up, both the internal and external tests attested to the versatility and stability of the proposed core lexicon. We expect this result to be applicable to future uses. The core lexicon should prove to be stable regardless of all sorts of textual variations.

4.2 Verification of the Applicability of the General Lexicon

As mentioned above, the general lexicon is constructed such that it will have comprehensive coverage of general texts not in a special domain. Thus, its goal is similar to that of the GB lexicon. Although, there are only 26,443 entries in our general lexicon, only 2/3 of the size of the GB lexicon (39,459 entries). However, our test will show that the disadvantage in size does not prevent the general lexicon form out-performing out-performing the GB lexicon.

(15) Textual Coverage of a Text by a Lexicon L

$\text{TextCov}_L \overset{def}{=}$ Number of tokens in the texts that are also L'2 entries /Total token number in the text

According to the above definition we can calculated the average textual coverage of all Sinica Corpus texts by our general lexicon to be 86.7619%; while the average textual coverage of the mush larger **GB** lexicon is only 83.3796%. The standard deviation of the coverage by the general lexicon is also almost 1% lower than that of the **GB** lexicon (3.9655% vs. 4.82408%). The lexicon coverage test also show similar results. In sum, we have attested that the general lexicon serves its purpose and our hybrid approach constructs a lexicon that out-performs one that is mainly corpus-based.

5. Conclusion

In this paper, we have proposed an approach to construct standard reference lexicons for NLP. This approach crucially depends on both corpora and lexical knowledge represented in human-compiled lexicons. In the process, we have also proposed formal principles to measure similarities between lexicons, as well as measures of coverage of a text by a lexicon. We use these formal measures to obtain data in support of our approach. We have also proposed a three level structure of standard lexicon, where the **Core Lexicon** will be the most versatile and most portable; the **General Lexicon** is less portable will be efficient and give comprehensive coverage for general applications; last, the reference lexicon is the open set reference that will contain as many words in the language as possible and will ideally allow users to extract their own special domain lexicons from; as well as to contribute their special domain entries to[2]. It is our hope that this first step towards a formal study of lexical selection principles as well as measurements for lexical coverage will point to a fertile

[2] The Lexicons are available under the following website:
http://rocling.iis.sinica.edu.tw

ground in computational lexicography, in addition to fulfilling its original goal of offering reliable data support for Chinese segmentation standard.

Acknowledgement: We would like to thank the helpful comments of professor C. C. Cheng and ROCLING reviewers, as well as CKIP colleques' help in preparing the data. Responsibilities for any remaining error is of course ours alone.

Reference

Armstrong-Warwick, S. 1995. Automated Lexical Resources in Europe: A Survey. In D.E. Walker, A. Zampolli, and N. Calzolari Eds. Automating the Lexicon. 397-403. Oxford: Oxford U. Press.

Chen, K.-j., C.-R. Huang, L.-P. Chang, and H.-L Hsu. 1996. SINICA CORPUS: Design Methodology for Balanced Corpora. In B.-S. Park and J.-B. Kim Eds. Language, Information, and Computation. Selected Papers from the 11th PACLIC. Seoul: Kynung Hee U.

Cheng, C. C. 1998. 從歷代經史子集研究人對語言詞彙的認知. 發表於人文計算研討會, Taipei, Academia Sinica..

Chinese Academy of Social Sciences. 1996. Xiandaihanyu Cidian [A Dictionary of Contemporary Chinese (Revised Edition)]. Beijing: Shangwu.

Chinese Knowledge Information Processing Group. 1996. ShouWen JieZi – A Study of Chinese Word Boundaries and Segmentation Standard for Information Processing [In Chinese]. CKIP Technical Report 96-01. Taipei: Academia Sinica.

---, 1995. The Grammatical Categories of Mandarin Chinese.[in Chinese] CKIP Technical Report 95-03. Taipei: Academia Sinica.

Huang, Chu-Ren, Keh-Jiann Chen, Feng-yi Chen, Wen-Jen Wei, and Lili Chang.

1997. The Design Criteria and Content of the Segmentation Standard for Chinese Information Processing [in Chinese]. Yuyan Wenzi Yingyong. 1997.1.92-100.

---, Keh-Jian Chen, Lili Chang and Feng-yi Chen. 1997. Segmentation Standard for Chinese Natural Language Processing. Computational Linguistics & Chinese Language Processing. 2.2.46-62.

---, Zhao-ming Gao, Claude C.C Shen, and Keh-jiann Chen. 1998. Towards a Sharable and Reusable Lexical List: The construction of a standard reference lexicon for Chinese NLP. Presented at the 1998 Pacific Neighborhood Consortion (PNC) Annual Meeting. To appear in the Proceedings. Taipei: Academia Sinica.

Lin, X.G., and C.J. Miao. 1997. Guifan+Cibiao yu Jinyen+Tongji. Yuyan Wenzi Yingyong. 1997.1.7-91.

Liu, Y., Q. Tan, and X. Shen. 1994. Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology.[in Chinese] Beijing: Qinghua U. Press.

Liu, Y., N. Liang, and Q. Tan. 1991. Lexical Selection Criteria for 'A Lexicon of Frequent Modern Mandarin Words for Information Processing'. Proceedings of the Tenth Anniversary of Chinese Information Society of China. 127-141.

Mei, J., Y. Zhu, Y. Gao, and H. Yin. 1983. Tongyici Cilin. Shanghai: Shangwu Press and Shanghai Dictionaries.

Nagao, M. 1994. A Methodology for the Construction of a Terminology Dictionary. In B.T.S. Atkins and A. Zampolli Eds. Computational Approaches to the Lexicon. 379-412. Oxford U. Press.

Sinclair, J. M. 1987. Ed. Looking Up-An account of the COBUILD Project in Lexical Computing. London: Collins. Sproat, R. 1992. Morphology and Computation.

Cambridge: MIT Press.

Sun, M.S., and L. Zhang. 1997. Renjibingcun, Zhiliangheyi –tantan Zhidinh xinxi chuliyong hanyu cibiao de celue. Yuyan Wenzi Yingyong. 1997.1.79-86.

Wang, M.-C., C.-R. Huang, and K.-j. Chen. 1995. The Idenfication and Classification of Unknown Words in Chinese: A N-gram- Based Approach. In A. Ishikawa and Y. Nitta Eds. The Proceedings of the 1994 Kyoto Conference. A Festschrift for Professor Akira Ikeya. 113-123. Tokyo: The Logico-Linguistics Society of Japan.

Zhang, Y. and X. Qi. 1997. The Statistics[s] and Analysis of Words Included in Several Chinese Dictionaries.[In Chinese] In L.W. Chen and Q. Yuan Eds. Language Engineering. 82-87. Beijing: Qinghua U. Press.

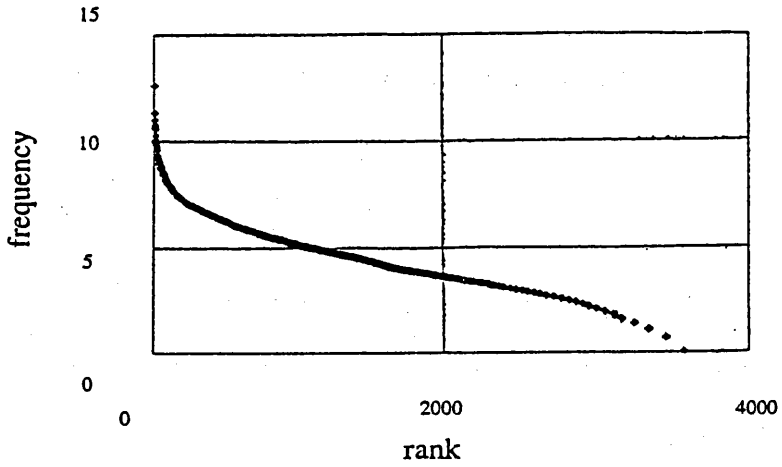

Diagram 1. CILIN entries frequency distribution in Sinica Corpus (Zipf's Law)
$Y = log\ f,\ X = rank$

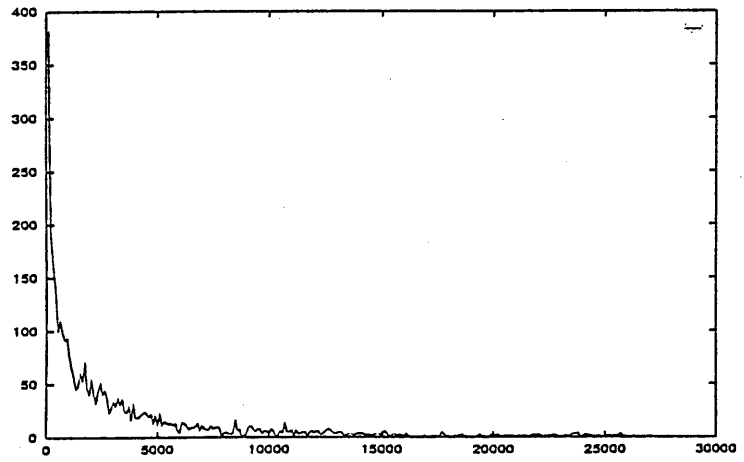

Diagram 2. CILIN entries distribution by frequency range in Sinica Corpus

(number of CILIN entries per every 1,000 rank interval)