The Role of Lexical Resources in CJK Natural Language Processing

Jack Halpern（春遍雀來）
The CJK Dictionary Institute (CJKI) (日中韓辭典研究所)
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001, Japan
jack@cjk.org

Abstract

The complexity of Chinese, Japanese and Korean (CJK) poses special challenges to developers of NLP tools, especially in the area of word segmentation (WS), information retrieval (IR), named entity extraction (NER), and machine translation (MT). These difficulties are exacerbated by the lack of truly comprehensive lexical resources, especially for proper nouns, and the lack of a standardized orthography, especially in Japanese. This paper summarizes some of the major linguistic issues in the development NLP applications that are heavily dependent on lexical resources, and discuses the central role such resources should play in enhancing the accuracy of NLP tools, especially for Chinese.

1 Introduction

Developers of CJK NLP tools face various challenges, some of the major ones being:

1. Identifying and processing the large number of orthographic variants in Japanese, and alternate character forms in CJK languages.
2. The lack of easily available comprehensive lexical resources, especially lexical databases, comparable to the major European languages.
3. The accurate conversion between Simplified and Traditional Chinese (Halpern and Kerman 1999).
4. The morphological complexity of Japanese and Korean.
5. Accurate word segmentation (Emerson 2000 and Yu et al. 2000) and disambiguating ambiguous segmentations strings (ASS) (Zhou and Yu 1994).
6. The difficulty of lexeme-based retrieval and CJK CLIR (Goto et al. 2001).
7. Miscellaneous technical requirements such as transcoding between multiple character sets and encodings, support for Unicode, and input method editors (Lunde 1999).
8. Chinese and Japanese proper nouns, which are extremely numerous and have many variants, are difficult to detect without a lexicon.
9. Automatic recognition of terms and their variants (Jacquemin 2001).

The various attempts to tackle these tasks by statistical and algorithmic methods (Kwok 1997) have had only limited success. An important motivation for such methodology has been the poor availability and great expense of acquiring and maintaining large-scale lexical databases.

This paper discusses how a lexicon-driven approach exploiting large-scale lexical databases can offer reliable solutions to some of the principal issues, based on over a decade of experience in building such databases for NLP applications.

2 Named Entity Extraction

Named Entity Recognition (NER) is useful in NLP applications such as question answering, machine translation and information extraction. A major difficulty in NER, and a strong motivation for using tools based on probabilistic methods, is that the compilation and maintenance of large entity databases is time consuming and expensive. The number of personal names and their variants (e.g. over a hundred ways to spell *Mohammed*) is probably in the billions. The number of place names is also large, though they are relatively stable compared with the names of organizations and products, which change frequently.

A small number of organizations, such as LAS and our institute, maintain databases of millions of proper nouns, but even such comprehensive databases cannot be kept fully up-to-date as countless new names are created daily. Various techniques have been used to automatically detect entities, one being the use of keywords or syntactic structures that co-occur with the entity, which we refer to as *named entity contextual clues* (NECC).

Table 1. Named Entity Contextual Clues

Headword	Reading	Example
センター	せんたー	国民生活センター
ホテル	ほてる	ホテルシオノ
駅	えき	朝霞駅
協会	きょうかい	日本ユニセフ協会

The above shows NECC table for Japanese personal names, which when used in conjunction with multilingual entity databases like the one below achieve high precision in entity recognition.

Table 2. Multilingual Database of Place Names

English	Japanese	Simplified Chinese	LO	Traditional Chinese	Korean
Azerbaijan	アゼルバイジャン	阿塞拜疆	L	亞塞拜然	아제르바이잔
Caracas	カラカス	加拉加斯	L	卡拉卡斯	카라카스
Cairo	カイロ	开罗	O	開羅	카이로
Chad	チャド	乍得	L	查德	차드
New Zealand	ニュージーランド	新西兰	L	紐西蘭	뉴질랜드
Seoul	ソウル	首尔	O	首爾	서울
Seoul	ソウル	汉城	O	漢城	서울
Yemen	イエメン	也门	L	葉門	예멘

Note how the lexemic pairs (“L” in the **LO** column) are not merely simplified and traditional *orthographic* (“O”) versions of each other, but independent lexemes equivalent to American *truck* and British *lorry.*

NER, especially of personal names and place names, is an area in which lexicon-driven methods have a clear advantage over probabilistic methods.

3 Linguistic Issues in Chinese

3.1 Processing Multiword Units

A major issue for Chinese segmentors is how to treat compound words and multiword lexical units (MWU), which are often decomposed into their components rather than treated as a single unit. For example, 录像带 *lùxiàngdài* 'video cassette' and 机器翻译 *jīqifānyì* 'machine translation' are not tagged as segments in Chinese Gigaword, the largest tagged Chinese corpus in existence, processed by the CKIP morphological analyzer (Ma 2003). Possible reasons for this include:

1. The lexicons used by Chinese segmentors are small-scale or incomplete. Our testing of various Chinese segmentors has shown that coverage of MWUs is often limited.
2. Chinese linguists disagree on the concept of wordhood in Chinese. Various theories such as the Lexical Integrity Hypothesis (Huang 1984) have been proposed. San Duanmu’s outstanding monograph (Duanmu 1998) on the subject clears up much of the confusion.
3. The "correct” segmentation can depend on the application, and there are various segmentation standards. For example a search engine user looking for 录像带 is not normally interested in 录像 'to videotape' and 带 'belt' per se, unless they are part of 录像带.

This last point is important enough to merit elaboration. A user searching for 中国人 *zhōngguórén* 'Chinese (person)' is *not* interested in 中国 'China', and vice-versa. A search for 中国 should *not* retrieve 中国人 as an instance of 中国. Exactly the same logic should apply to 机器翻译, so that a search for that keyword should only retrieve documents containing that string in its entirety. Yet performing a Google seach on 机器翻译 in normal mode gave some 2.3 million hits, hudreds of thousands of which had zero occurrences of 机器翻译 but numerous occurrences of unrelated words like 机器人 'robot', which the user is not interested in.

This is equivalent to saying that *headwaiter* should not be considered an instance of *waiter*, which is indeed how Google behaves. More to the point, English space-delimited lexemes like *high school* are not instances of the adjective *high*. As shown in Halpern (2000b), "the degree of solidity often has nothing to do with the status of a string as a lexeme. *School bus* is just as legitimate a lexeme as is *headwaiter* or *word-processor*. The presence or absence of spaces or hyphens, that is, the orthography, does not determine the lexemic status of a string."

In a similar manner, it is perfectly legitimate to consider Chinese MWUs like those shown below as indivisible units for most applications, especially information retrieval and machine translation.

丝绸之路	*sīchóuzhīlù*	silk road
机器翻译	*jīqifānyì*	machine translation
爱国主义	*àiguózhǔyì*	patriotism
录像带	*lùxiàngdài*	video cassette
新西兰	*Xīnxīlán*	New Zealand
临阵磨枪	*línzhènmóqiāng*	start to prepare at the last moment

One could argue that 机器翻译 is compositional and therefore should be considered "two words." Whether we count it as one or two "words" is not really relevant – what matters is that it is *one lexeme* (smallest distinctive units associating meaning with form). On the other extreme, it is clear that idiomatic expressions like 临阵磨枪, literally "sharpen one's spear before going to battle," meaning 'start to prepare at the last moment,' are indivisible units.

Predicting compositionality is not trivial and often impossible. For many purposes, the only practical solution is to consider all lexemes as indivisible. Nonetheless, currently even the most advanced segmentors fail to identify such lexemes and missegment them into their constituents, no doubt because they are not registered in the lexicon. This is an area in which expanded lexical resources can significantly improve segmentation accuracy.

In conclusion, lexical items like 机器翻译 'machine translation' represent stand-alone, well-defined concepts and should be treated as single units. The fact that in English *machineless* is spelled solid and *machine translation* is not is an historical accident of orthography unrelated to the fundamental fact that both are full-fledged lexemes each of which represents an indivisible, indepenedent concept. The same logic applies to 机器翻译, which is a full-fledged lexeme that should not be decomposed.

3.2 Multilevel Segmentation

Chinese MWUs can consist of nested components that can be segmented in different ways for different levels to satisfy the requirements of different segmentation standards. The example below shows how 北京日本人学校 *Bĕijīng Rìbĕnrén Xuéxiào* 'Beijing School for Japanese (nationals)' can be segmented on five different levels.

1. 北京日本人学校 multiword lexemic
2. 北京+日本人+学校 lexemic
3. 北京+日本+人+学校 sublexemic
4. 北京 + [日本 + 人] [学+校] morphemic
5. [北+京] [日+本+人] [学+校] submorphemic

A more advanced and expensive solution is to store presegmented MWUs in the lexicon, or even to store nesting delimiters as shown above, giving the user the option to select the desired segmentation level.

This problem is especially obvious in the case neologisms. Of course no lexical database can expect to keep up with the latest neologisms, and even the first edition of Chinese Gigaword does not yet have 博客 *bókè* 'blog'. Here are some examples of MWU neologisms, some of which are not (at least bilingually), compositional but fully qualify as lexemes.

仓储式连锁店	*cāngchǔshìliánsuǒdiàn*	warehouse club
电脑迷	*diànnǎomí*	cyberphile
电子商务	*diànzǐshāngwù*	e-commerce
追车族	*zhuīchēzú*	auto fan

3.3 Chinese-to-Chinese Conversion (C2C)

Numerous Chinese characters underwent drastic simplifications in the postwar period. Chinese written in these simplified forms is called Simplified Chinese (SC). Taiwan, Hong Kong, and most overseas Chinese continue to use the old, complex forms, referred to as Traditional Chinese (TC). Contrary to popular perception, the process of accurately converting SC to/from TC is full of complexities and pitfalls. The linguistic issues are discussed in Halpern and Kerman (1999), while technical issues are described in Lunde (1999). The conversion can be implemented on three levels in increasing order of sophistication:

1. Code Conversion. The easiest, but most unreliable, way to perform C2C is to transcode by using a one-to-one mapping table. Because of the numerous one-to-many ambiguities, as shown below, the rate of conversion failure is unacceptably high.

Table 3. Code Conversion

SC	TC1	TC2	TC3	TC4	Remarks
门	們				one-to-one
汤	湯				one-to-one
发	發	髮			one-to-many
暗	暗	闇			one-to-many
干	幹	乾	干	榦	one-to-many

2. Orthographic Conversion. The next level of sophistication is to convert orthographic units, rather than codepoints. That is, meaningful linguistic units, equivalent to lexemes, with the important difference that the TC is the traditional version of the SC on a character form level. While code conversion is ambiguous, orthographic conversion gives much better results because the orthographic mapping tables enable conversion on the lexeme level, as shown below.

Table 4. Orthographic Conversion

English	SC	TC1	TC2	Incorrect
Telephone	电话	電話		
Dry	干燥	乾燥		干燥 幹燥 榦燥
	阴干	陰乾	陰干	

As can be seen, the ambiguities inherent in code conversion are resolved by using orthographic mapping tables, which avoids false conversions such as shown in the **Incorrect** column. Because of segmentation ambiguities, such conversion must be done with a segmentor that can break the text stream into meaningful units (Emerson 2000).

An extra complication, among various others, is that in some lexemes have one-to-many orthographic mappings, *all* of which are correct. For example, SC 阴干 correctly maps to both TC 陰乾 'dry in the shade' and TC 陰干 'the five even numbers'. Well designed orthographic mapping tables must take such anomalies into account.

3. Lexemic Conversion. The most sophisticated form of C2C conversion is called *lexemic conversion,* which maps SC and TC lexemes that are semantically, not orthographically, equivalent. For example, SC 信息 *xìnxī* 'information' is converted into the semantically equivalent TC 資訊 *zīxùn*. This is similar to the difference between British *pavement* and American *sidewalk.* Tsou (2000) has demonstrated that there are numerous lexemic differences between SC and TC, especially in technical terms and proper nouns, e.g. there are more than 10 variants for *Osama bin Laden.*

Table 5. Lexemic Conversion

English	SC	Taiwan TC	HK TC	Incorrect TC
Software	软件	軟體	軟件	軟件
Taxi	出租汽车	計程車	的士	出租汽車
Osama Bin Laden	奥萨马本拉登	奧薩瑪賓拉登	奧薩瑪賓拉丹	奧薩馬本拉登
Oahu	瓦胡岛	歐胡島		瓦胡島

3.4 Traditional Chinese Variants

Traditional Chinese has numerous variant character forms, leading to much confusion. Disambiguating these variants can be done by using mapping tables such as the one shown below. If such a table is carefully constructed by limiting it to cases of 100% semantic interchangeability for polysemes, it is easy to normalize a TC text by trivially replacing variants by their standardized forms. For this to work, all relevant components, such as MT dictionaries, search engine indexes and the related documents should be normalized. An extra complication is that Taiwanese and Hong Kong variants are sometimes different (Tsou 2000).

Table 6. TC Variants

Var. 1	Var. 2	English	Comment
裏	裡	Inside	100% interchangeable
著	着	Particle	variant 2 not in Big5
沉	沈	sink; surname	partially interchangeable

4 Orthographic Variation in Japanese

4.1 Highly Irregular Orthography

The Japanese orthography is highly irregular, significantly more so than any other major language, including Chinese. A major factor is the complex interaction of the four scripts used to write Japanese, e.g. kanji, hiragana, katakana, and the Latin alphabet, resulting in countless words that can be written in a variety of often unpredictable ways, and the lack of a standardized orthography. For example, *toriatsukai* 'handling' can be written in six ways: 取り扱い, 取扱い, 取扱, とり扱い, 取りあつかい, とりあつかい.

An example of how difficult Japanese IR can be is the proverbial 'A hen that lays golden eggs.' The "standard" orthography would be 金の卵を産む鶏 *Kin no tamago wo umu niwatori*. In reality, tamago 'egg' has four variants (卵, 玉子, たまご, タマゴ), niwatori 'chicken' three (鶏, にわとり, ニワトリ) and umu 'to lay' two (産む, 生む), which expands to 24 permutations like 金の卵を生むニワトリ, 金の玉子を産む鶏 etc. As can be easily verified by searching the web, these variants occur frequently.

Linguistic tools that perform segmenation, MT, entity extraction and the like must identify and/or normalize such variants to perform dictionary lookup. Below is a brief discussion of what kind of variation occurs and how such normalization can be achieved.

4.2 Okurigana Variants

One of the most common types of orthographic variation in Japanese occurs in kana endings, called *okurigana*, that are attached to a kanji stem. For example, *okonau* 'perform' can be written 行う or 行なう, whereas *toriatsukai* can be written in the six ways shown above. Okurigana variants are numerous and unpredictable. Identifying them must play a major role in Japanese orthographic normalization. Although it is possible to create a dictionary of okurigana variants algorithmically, the resulting lexicon would be huge and may create numerous false positives not semantically interchangeable. The most effective solution is a database of okurigana variants, such as the one shown below:

Table 7. Okurigana Variants

HEADWORD	READING	NORMALIZED
書き著す	かきあらわす	書き著す
書き著わす	かきあらわす	書き著す
書著す	かきあらわす	書き著す
書著わす	かきあらわす	書き著す

Since Japanese is highly agglutinative and verbs can have numerous inflected forms, a table such as the above must be used in conjunction with a morphological analyser that can do accurate stemming, i.e. be capable of recognizing that 書き著しませんでした is the polite form of the canonical form 書き著す.

4.3 Cross-Script Orthographic Variation

Variation across the four scripts in Japanese is common and unpredictable, so that the same word can be written in any of several scripts, or even as a hybrid of multiple scripts, as shown below:

Table 8. Cross-Script Variation

Kanji	Hiragana	katakana	Latin	Hybrid	Gloss
人参	にんじん	ニンジン			carrot
		オープン	OPEN		open
硫黄		イオウ			sulfur
		ワイシャツ		Yシャツ	shirt
皮膚		ヒフ		皮フ	skin

Cross-script variation can have a major consequences for recall, as can be seen from the table below.

Table 9: Hit Distribution for 人参 'carrot' *ninjin*

ID	Keyword	Normalized	Google Hits	Formula
A	人参	人参	67,500	A＋α_{1}
B	にんじん	人参	66,200	B＋α_{2}
C	ニンジン	人参	58,000	C＋α_{3}

Using the ID above to represent the number of Google hits, this gives a total of A＋B＋C＋α_{123} = 191,700. α is a coincidental occurrence factor, such as in '100 人参加, in which '人参' is unrelated to the 'carrot' sense. The formulae for calculating the above are as follows.

Unnormalized recall:

$$\frac{C}{A+B+C+\alpha_{123}}=\frac{58,\ 000}{191,\ 700}\ (\approx 30\%)$$

Normalized recall:

$$\frac{A+B+C}{A+B+C+\ \alpha_{123}}=\frac{191,\ 700}{191,\ 700}\ (\approx 100\%)$$

Unnormalized precision:

$$\frac{C}{C+\alpha_{3}}=\frac{58,\ 000}{58,\ 000}\ (\approx 100\%)$$

Normalized precision:

$$\frac{C}{A+B+C+\alpha_{123}}=\frac{191,\ 700}{191,\ 700}\ (\approx 100\%)$$

人参 'carrot' illustrates how serious a problem cross-orthographic variants can be. If orthographic normalization is not implemented to ensure that all variants are indexed on a standardized form like 人参, recall is only 30%; if it is, there is a dramatic improvement and it goes up to nearly 100%, without any loss in precision, which hovers at 100%.

4.4 Kana Variants

A sharp increase in the use of katakana in recent years is a major annoyance to NLP applications because katakana orthography is often irregular; it is quite common for the same word to be written in multiple, unpredictable ways. Although hiragana orthography is generally regular, a small number of irregularities persist. Some of the major types of kana variation are shown in the table below.

Table 10. Kana Variants

Type	English	Standard	Variants
Macron	computer	コンピュータ	コンピューター
Long vowels	maid	メード	メイド
Multiple kana	team	チーム	ティーム
Traditional	big	おおきい	おうきい
づ vs. ず	continue	つづく	つずく

The above is only a brief introduction to the most important types of kana variation. Though attempts at algorithmic solutions have been made by some NLP research laboratories (Brill 2001), the most practical solution is to use a katakana normalization table, such as the one shown below, as is being done by Yahoo! Japan and other major portals.

Table 11. Kana Variants

HEADWORD	NORMALIZED	English
アーキテクチャ	アーキテクチャー	architecture
アーキテクチャー	アーキテクチャー	architecture
アーキテクチュア	アーキテクチャー	architecture

4.5 Miscellaneous Variants

There are various other types of orthographic variants in Japanese, described Halpern (2000a). To mention some, kanji even in contemporary Japanese often have variants, such as 才 for 歳 and 巾 for 幅 and traditional forms such as 發 for 発. In addition, the large number of *kun* homophones and their variable orthography are often close or even identical in meaning, i.e., *noboru* means 'go up' when written 上る but 'climb' when written 登る, so that great care must be taken in the normalzation process so as to assure semantic interchangeability.

4.6 Lexicon-driven Normalization

Leaving statistical methods aside, lexcion-driven normalization of Japanese orthographic variants can be achieved by using an orthographic mapping table such as the one shown below, using various techniques such as:

1. Convert variants to a standardized form for indexing.
2. Normalize queries for dictionary lookup.
3. Normalize all source documents.
4. Identify forms as members of a variant group.

Table 12. Orthographic Normalization Table

HEADWORD	READING	NORMALIZED
空き缶	あきかん	空き缶
空缶	あきかん	空き缶
明き罐	あきかん	空き缶
あき缶	あきかん	空き缶
あき罐	あきかん	空き缶
空きかん	あきかん	空き缶
空きカン	あきかん	空き缶
空き罐	あきかん	空き缶
空罐	あきかん	空き缶
空き鑵	あきかん	空き缶
空鑵	あきかん	空き缶

Other possibilities for normalization include advanced applications such as domain-specific synonym expansion, requiring Japanese thesauri based on domain ontologies, as is done by a select number of companies like Wand and Convera who build sophisticated Japanese IR systems.

5 Orthographic Variation in Korean

Modern Korean has is a significant amount of orthographic variation, though far less than in Japanese. Combined with the morphological complexity of the language, this poses various challenges to developers of NLP tools. The issues are similar to Japanese in principle but differ in detail.

Briefly, Korean has variant hangul spellings in the writing of loanwords, such as 케이크 *keikeu* and 케잌 *keik* for 'cake', and in the writing of non-Korean personal names, such as 클린턴 *keulrinteon* and 클린톤 *keulrinton* for 'Cinton'. In addition, simiar to Japanese but on a smaller scale, Korean is written in a mixture of hangul, Chinese characters and the Latin alphabet. For example, 'shirt' can be written 와이셔츠 *wai-syeacheu* or Y 셔츠 *wai-syeacheu*, whereas 'one o'clock' can written as 한시 *hansi*, 1 시 *hansi* or 一時 *hansi*. Another issue is the differences between South and North Korea spellings, such as N.K. 오사까 *osakka* vs. S.K. 오사카 *osaka* for 'Osaka', and the old (pre-1988) orthography versus the new, i.e. modern 일군 'worker' (*ilgun*) used to be written 일꾼 (*ilkkun*).

Lexical databases, such as normaization tables similar to the ones shown above for Japanese, are the only practical solution to identifying such variants, as they are in principle unpredictable.

6 The Role of Lexical Databases

Because of the irregular orthography of CJK languages, procedures such as orthographic normalization cannot be based on statistical and probabilistic methods (e.g. bigramming) alone, not to speak of pure algorithmic methods. Many attempts have been made along these lines, as for example Brill (2001) and Goto et al. (2001), with some claiming performance equivalent to lexicon-driven methods, while Kwok (1997) reports good results with only a small lexicon and simple segmentor.

Emerson (2000) and others have reported that a robust morphological analyzer capable of processing lexemes, rather than bigrams or n-grams, must be supported by a large-scale computational lexicon. This experience is shared by many of the world's major portals and MT developers, who make extensive use of lexical databases.

Unlike in the past, disk storage is no longer a major issue. Many researchers and developers, such as Prof. Franz Guenthner of the University of Munich, have come to realize that “language is in the data,” and “the data is in the dictionary,” even to the point of compiling full-form dictionaries with millions of entries rather than rely on statistical methods, such as Meaningful Machines who use a full form dictionary containing millions of entries in developing a human quality Spanish-to-English MT system.

Our institute, which specializes in CJK and Arabic computational lexicography, is engaged in an ongoing research and development effort to compile CJK and Arabic lexical databases (currently about seven million entries), with special emphasis on proper nouns, orthographic normalization, and C2C. These resources are being subjected to heavy industrial use under real-world conditions, and the feedback thereof is being used to further expand these databases and to enhance the effectiveness of the NLP tools based on them.

Conclusions

Performing such tasks as orthographic normalization and named entity extraction accurately is beyond the ability of statistical methods alone, not to speak of C2C conversion and morphological analysis. Because of the irregular orthography of the CJK writing systems, information retrieval requires not only sophisticated tools such as morphological analysers, but also lexical databases fine-tuned to the needs of NLP applications. The building of large-scale lexicons based on corpora consisting of even billions of words has come of age. Since lexicon-driven techniques have proven their effectiveness, there is no need to overly rely on probabilistic methods. Comprehensive, up-to-date lexical resources are the key to achieving major enhancements in NLP technology.

References

Brill, E. and Kacmarick, G. and Brocket, C. (2001) *Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs*. Microsoft Research, Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan.

Duanmu, San (1998) *Wordhood in Chinese.* In “New Approaches to Chinese Word Formation”, Mouton Degruyter, Berlin and New York.

Emerson, T. (2000) *Segmenting Chinese in Unicode. Proc. of the 16th International Unicode Conference*, Amsterdam

Goto, I., Uratani, N. and Ehara T. (2001) *Cross-Language Information Retrieval of Proper Nouns using Context Information*. NHK Science and Technical Research Laboratories. Proc. of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan

Huang, James C. (1984) *Phrase Structure, Lexical Integrity, and Chinese Compounds,* Journal of the Chinese Teachers Language Association, 19.2: 53-78

Jacquemin, C. (2001) *Spotting and Discovering Terms through Natural Language Processing*. The MIT Press, Cambridge, MA

Halpern, J. and Kerman J. (1999) *The Pitfalls and Complexities of Chinese to Chinese Conversion*. Proc. of the Fourteenth International Unicode Conference in Cambridge, MA.

Halpern, J. (2000a) *The Challenges of Intelligent Japanese Searching*. Working paper (www.cjk.org/cjk/joa/joapaper.htm), The CJK Dictionary Institute, Saitama, Japan.

Halpern, J. (2000b) *Is English Segmentation Trivial?*. Working paper, (www.cjk.org/cjk/reference/engmorph.htm) The CJK Dictionary Institute, Saitama, Japan.

Kwok, K.L. (1997) *Lexicon Effects on Chinese Information Retrieval*. Proc. of 2nd Conf. on Empirical Methods in NLP. ACL. pp.141-8.

Lunde, Ken (1999) *CJKV Information Processing*. O’Reilly & Associates, Sebastopol, CA.

Yu, Shiwen, Zhu, Xue-feng and Wang, Hui (2000) *New Progress of the Grammatical Knowledge-base of Contemporary Chinese*. Journal of Chinese Information Processing, Institute of Computational Linguistics, Peking University, Vol.15 No.1.

Ma, Wei-yun and Chen, Keh-Jiann (2003) *Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff,* Proceedings of the Second SIGHAN Workshop on Chinese Language Processingpp. 168-171 Sapporo, Japan

Yu, Shiwen, Zhu, Xue-feng and Wang, Hui (2000) *New Progress of the Grammatical Knowledge-base of Contemporary Chinese*. Journal of Chinese Information Processing, Institute of Computational Linguistics, Peking University, Vol.15 No.1.

Tsou, B.K., Tsoi, W.F., Lai, T.B.Y. Hu, J., and Chan S.W.K. (2000) *LIVAC, a Chinese synchronous corpus, and some applications*. In "2000 International Conference on Chinese Language ComputingICCLC2000", Chicago

Zhou, Qiang. and Yu, Shiwen (1994) *Blending Segmentation with Tagging in Chinese Language Corpus Processing,* 15th International Conference on Computational Linguistics (COLING 1994)