Project, and the data set made available for the SemEval 2007 task on the word sense disambiguation of prepositions (Litkowski, 2007), providing the means for more detailed analysis of preposition semantics. In 2009, a special issue of Computational Linguistics, 35(2), is devoted to preposition research. It gave a survey of preposition applications. Baldwin(2009) pointed out that prepositions had led a mixed existence in computational linguistics and related fields, particularly in the context of applications. He mainly summarized the syntax and semantic developments in prepositions, including PP attachment, prepositional multiword expression, lexical semantic resources for prepositions, automatic classification of preposition sense, and so on. Besides, Jørgensen(2009) presented a formal semantic analysis of preposition semantics, and its possible application in Norwegian-English machine translation. Kelleher(2009) proposed computational models of topological and projective spatial prepositions for use in a visually situated dialogue system. O'Hara(2009) discussed on how semantic role resources could be exploited for preposition disambiguation. Girju(2009) leveraged translation data to improve the accuracy of compound noun interpretation, based on the observation that the choice of preposition in Romance languages is often indicative of the semantics of the compound noun.

But the special issue of Computational Linguistics did not refer to any development in Chinese information processing. The function words are very important for Chinese analysis and generation. Even the style of an author's writing can be identified only by the function words usages in his or her works. Setiawan(2009) introduced topological ordering of function words in hierarchical phrase-based translation, and got significant gains in his Chinese-English and Arabic-English experiments. In fact, there have been much research on Chinese function words and the relative applications. For example, the auxiliary 的, the top one frequent word in Chinese text, has the most various usages in different context and means a lot in Chinese understanding. Linking constructions involving 的 are ubiquitous in Chinese, and can be translated into English in many different ways. Chang(2009) pointed out that 的 is a major source of word order error, and explored how to get more information about the syntactic, semantic, and discourse context of 的 usages to facilitate producing an appropriate English translation strategy. He got significant BLEU point gains on MT02 (+1.24), MT03 (+0.88) and MT05 (+1.49) on a phrased based system. Follow this work, DU(2010) did further experiments on his proposed model, and outperformed the baseline systems by 6.42% and 3.08% relative points in terms of the BLEU score on PB-SMT and hierarchical phrase-based MT respectively.

Therefore, function words usages identification is a key factor in Chinese information processing. However, most of above researches have focused on several individuals, and have not yet formed a complete dictionary or knowledge base on Chinese function word usages.

3 CFKB

The construction of CFKB was kicked off in 2003, and the "trinity" design concept of CFKB was prompted by Yu(Yu, 2003). The basic frame of CFKB was discussed in Liu(2004), including the attributes of the machine dictionary, the proper-scope of corpus and rule base. Peng(2006) studied grammatical functions of Chinese prepositions, and gave the preliminary machine dictionary and rule base of Chinese prepositions. Zan(2007) and Hao(2007) gave a formal descriptions on Chinese adverb, and built a tentative machine dictionary and rule base of Chinese adverbs.

Complying with the trinity design concept(Yu, 2003), CFKB includes Chinese function word usage dictionary, Chinese function word usage rule base, and Chinese function word usage corpus.

3.1 Chinese function word usage dictionary

CFKB includes all the function words from GKB (Yu, 1998), and gives more detail description on the words usage. For example, 才 has only one adverb ID in GKB. But in contemporary Chinese real texts, 才 would has 5 senses with 9 usages in total when it is in different context (Lv, 1980). So there are as many as 9 IDs for adverb才 in CFKB. They are <d_cai2_1a>, <d_cai2_1b>, <d_cai2_2a>, <d_cai2_2b>, <d_cai2_3>, <d_cai2_4>, <d_cai2_5a>, <d_cai2_5b>, and <d_cai2_5c>.

Each usage has a unique ID in the function usage dictionary. The coding principles of ID in CFKB is based on Lv (1980). In general, the ID is in the frame of <x_y_nz>, where "x" represents the word's POS, "y" represents the word's PINYIN (Chinese pronouncing notation), "n" represents the sense number, and "z" represents the usage in sense "z". If there are multiple different function words sharing the same pronunciation, we will use the frame <x_y_tm_nz>, where "t" shows it is a homophone with other word, and "m" represents its sequent number. The amount of ID for a function word depends on the word's usages. There are considerable variations among different function words. The more frequent the words are, the more usages they will have in CFKB. There are many function words which have only 1 sense with 1 usage in total, but for the words with high frequency in text will have much more senses or usages. For example, proposition 被 has 1 sense with 8 usages in total, adverb 就 has 7 senses with 21 usages in total, and auxiliary 的 has 11 senses with 39 usages in total. This just proves that function word has strong properties to study finely.

Besides GBK, CFKB also referred to Lv(1980), Zhang(2001), "Contemporary Chinese Dictionary"(5th edition), and the segmentation and POS corpus of People's Daily, Jan 1998, and Jan to Jun 2000. With further cognition, the function words senses and usages in CFKB are continually changing or adjusting slightly. Table 1 shows the present distribution of Chinese function words' usages in CFKB.

Usages No / Words No. / POS	1	2	3	4	5	6	7	8	9	10	Above 10	Words in total	Usages in total
Adverbs	1214	179	84	38	21	12	4	3	2	1	8	1566	2356
Preposition	66	30	23	7	4	5	7	0	1	1	2	141	331
Conjunction	156	50	55	24	16	7	4	0	1	2	0	315	696
Auxiliary	30	4	3	1	0	1	2	0	0	0	1	45	144
Modality	30	7	7	4	2	0	1	4	0	0	2	58	169

Table 1 The distribution of Chinese function words' usages in CFKB.

CFKB adopts the relational database form to describe each function word usage in detail with usage-attribute feature. CFKB includes all the features from GKB(Yu, 1998), and many of them are replenished or adjusted for each usage of the function words, for example, the sense, the usage, the example sentence, subclass information, and so on. Besides the features inherited from GKB, CFKB increases usage ID and other six usage-attributes for each usage of the function words. The usage attributes are all about the special collocation or context of the observed function word, and will be adopted in their usage rules. They are as follows.

- the features of the first word in a sentence (short for F);
- the features of words to the left of the function word in a sentence (short for M);
- the features of words close to the left of the function word in a sentence (short for L);
- the feature of words close to the right of the function word in a sentence (short for R);
- the features of words to the right of the function word in a sentence (short for N);
- the features of end word in a sentence (short for E).

3.2 Chinese function word usage rule base

Based on Chinese function word usage dictionary and the statistical laws of function words' occurrence in People's Daily, we have distilled feasible criterions and have written the usage rules in Backus Normal Form (BNF). The usage rules for each function word are ordered for higher precision of automatic identification, not ordered by ID sequence. The meta symbols, "→", "|", "*", "( )" and "[ ]", are in the general means of BNF. The symbol "#" we introduced is to express any word string except nothing. And the symbol "~" represents the observed function word itself.

To describe the usage rules, we utilized the six usage features such as F, M, L, R, N, and E, which are explained in section 3.1. In Chinese function word usage rule base, each Chinese function word starts with "$", while each usage start with "@". If there are multiple rules in one usage, the following rules start with "^". Here are some usage rule examples,

$特别
@<d_te4bie2_3a>→FR ^F→~ ^R→[是]*n
@<d_te4bie2_3b>→FR ^F→~ ^R→v
@<d_te4bie2_2>→N ^N→v
@<d_te4bie2_1>→N ^N→v|a

$那么
@<c_na4me5_t1_1b>→M ^M→？[”]
@<c_na4me5_t1_3b>→L ^L→。
@<c_na4me5_t1_3a>→L ^L→既然#，
@<c_na4me5_t1_1a>→L ^L→(如果|要是)#，
@<c_na4me5_t1_2>→L ^L→(如果|如果说)#，

In the above examples, the "特别" is a Chinese adverb and has 3 senses with 4 usages in total. Among them, the first sense, < d_te4bie2_1>, means "very". The second sense, <d_te4bie2_2> , means "specially". The third sense, <d_te4bie2_3a> and < d_te4bie2_3b>, means "particularly". The difference between two usages of the third sense is that in <d_te4bie2_3a>, "特别" modifies a noun, whereas in < d_te4bie2_3b>, "特别" modifies a verb.

Each usage rule defines its special context features. For example, the usage <d_te4bie2_1> of "特别" is defined as there should be a verb (short for v) or an adjective in the right context of "特别" in observed sentence. The usage <d_te2bie4_3a> is defined as two conditions, F and R, to be satisfied simultaneously. Here F is defined as that the first word should be "特别" itself in observed sentence, while R is defined as there should be a structure "[是]*n" in the close right context in the observed sentence. The "[是]*n" means the word "是" is optional. And then, "*" means any word, including nothing. That is, there can be any word between "是" and the noun (short for n).

The "那么" is a Chinese conjunction and has 3 senses with 5 usages in total. The first sense, <c_na4me5_t1_1a> and <c_na4me5_t1_1b>, expresses a relation of hypothesis. The second sense, <c_na4me5_t1_2>, expresses a relation of inference. The third sense, <c_na4me5_t1_3a> and <c_na4me5_t1_3b>, expresses a relation of deduction. You can get the usages' difference from the description of rule. And you must have found the rules ID of "那么" include "t1". This shows there is another function word in the same pronunciation with "那么", such as "那末".

Up to now, we have finished the descriptions of all usage rules and constructed the initial Chinese function word usage rule base. Further works include adjusting or modifying the rules to express their context more appropriately, and to get higher precision for rule based automatic identification of Chinese function word usages.

3.3 Chinese function word usage corpus

The Chinese function word usage corpus is an important part of the trinity CFKB. For about recent seven years, we have devoted to the Chinese function word usage annotating on the segmentation and POS corpus for 7 months of People's Daily, Jan, 1998, and Jan to Jun, 2000.

The total of words in each month of People's Daily corpus is about 1.2 million. Among them, as for the occurrence of function words, there are about 50,000 adverbs, over 40,000 prepositions, nearly 30,000 conjunctions, nearly 80,000 auxiliaries and less than 2,000 modalities. And for each particular function word, the situation is quite different among each other. For example, the auxiliary 的 has the highest frequency with about 40,000 occurrences in each month corpus. However, there are many other function words which do not occur at all.

To annotate Chinese function words usage on such a large size corpus is a difficult and time-consuming task. First, based on Chinese function word usage dictionary and rule base, we got the elementary tagged result via a rule based automatic identification system of function word usages. Because of the flexible usages of the function words in real text, there must be many errors in the tags in this elementary result to be corrected. Then two annotators checked

the corpus double blindly and a committee decided those inconsistent annotations. To improve the efficiency and correctness of annotators' work, we have developed an aided tool which integrates the three part of CFKB. Figure 1 shows the user interface of the aided tool.

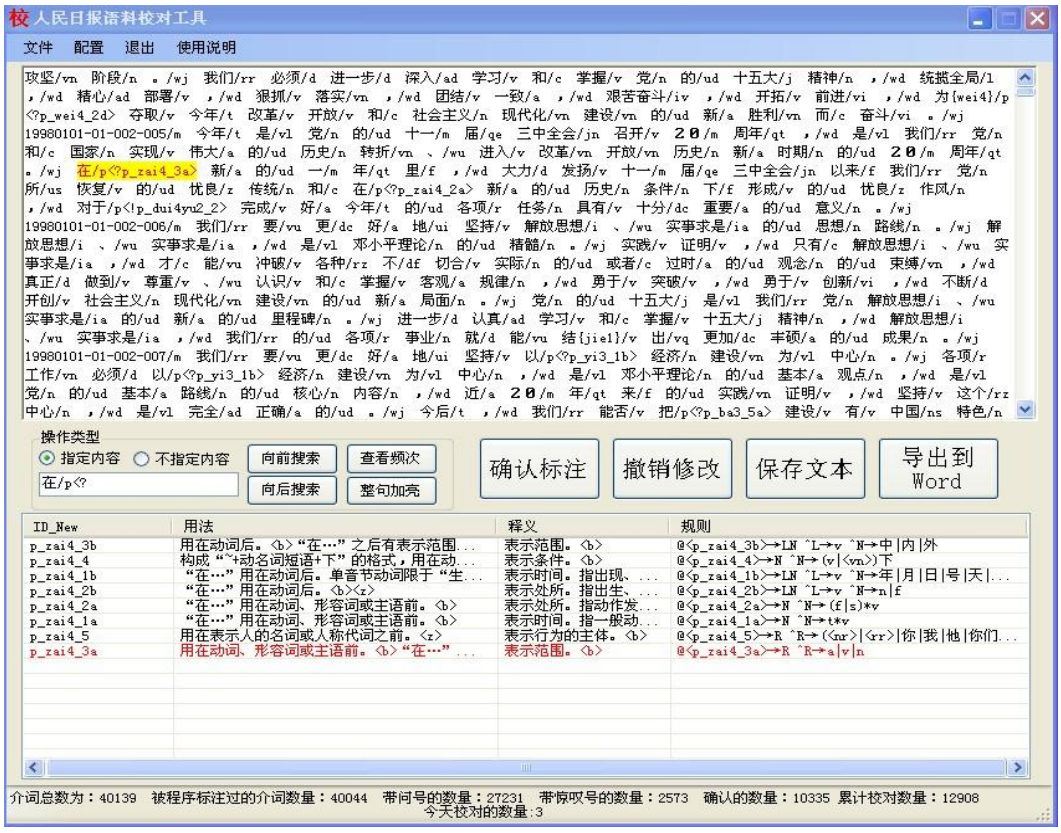

Figure 1 The interface of aided tools for checking Chinese function word usages corpus

Here are some examples of the Chinese function word usage corpus.

- 我们/r 要/vu **更/dc<d_geng4_1a>** 好/a **地/u< u_de5_t3>** 坚持/v "/wyz 两手抓/l 、/wu 两手/mq **都/d<d_dou1_1>** 要/vu 硬/a "/wyy **的/u< u_de5_t2_1g >** 方针/n 。/wj

- 北方/f 省/n **的/u< u_de5_t2_1a>** 铬铁矿/n 企业/n **在/p< p_zai4_1b >** 短/a 时间/n 内/f **就/d<d_jiu4_4a>** 开始/v 运作/v ，/wd **仅/d<d_jin3_t1_1aa>** 这/rz 一/m 项目/n 投资/n **就/d<d_jiu4_6b>** 逾/Vg 7 0 0 0 万/m 美元/qd 。/wj

Up to now, we have finished all annotating work of adverbs, prepositions, conjunctions, modalities, and most of auxiliaries. With this Chinese function word usage annotated corpus, we can supplement the frequency information into the Chinese function word usage dictionary, and can also utilize the frequency information to improve the performance of Chinese function word usage rule base. We believe that the Chinese function word usage corpus will become an important data for the research of Chinese parsing or generation, and other Chinese related natural language processing applications, for example, machine translation, information extraction, question answer, and so on.

3.4 Automatic identification algorithms of Chinese function words usages

Two kinds of usage automatic identification algorithms were developed. One was based on usage rules, and the other was based on statistical model.

As for rule based method, Liu(2008) developed the initial automatic identification system for Chinese adverb usages with the average precision at about 70% - 80%. Yuan(2010) rewrote the adverb usages' automatic identification system and expanded it for all kinds of Chinese function words. The system precision depends on the rules in rule base. The precisions are quite different among various POS, and among the individual words. The precision of several words' usage identification based on rule were lower than 30%, and several others' were even up to 100%. This is consistent with the common cognition that the Chinese function words have strong individual characteristics with distinct usages. The usage rules must be carefully studied one by one, and would be improved gradually. Zhou(2010) gave some efficient improvement on Chinese modality 了.

It is well known, only rule based methods are not enough for natural language processing, especially for so difficult Chinese function word usage's identification. After manually checking and adjusting the tagged results of rule based system, we got the gold standard Chinese function words usage corpus. With this training data, statistical models were adopted in the automatic identification of Chinese function word usages. For example, the usages' automatic identification of adverb 才(Zan, 2009) and several common adverbs (Zan, 2010) were tried via the model of ME (Maximum Entropy), CRF (Conditional Random Fields), and SVM (Support Vector Machine). They got higher average precision by appropriate feature selection for most experimental adverbs. Table 2 shows some experiment results of rule based method and the statistical based methods (Zan, 2010).

Method / Adverb	Rule-based	CRF	ME	SVM
bian/便	0.409	0.459	0.453	0.876
fenbie/分别	0.506	0.673	0.679	0.905
Jiu/就	0.339	0.776	0.608	0.59
tebie/特别	0.697	0.783	0.652	0.932
yi/已	0.511	0.91	0.71	0.974
shifen/十分	0.712	0.95	0.865	0.993
xianhou/先后	0.963	0.575	0.59	0.846
average precision	**0.55**	**0.729**	**0.66**	**0.885**

Table 2 The experiment results of rule based method and the statistical based methods.

4 Conclusion and Future Works

Chinese function words are very important in text semantic understanding and grammatical analysis. In this paper, we reviewed the research results and applications of Chinese function words, and put forward the necessity to study the Chinese function word usages systematically. We introduced all the three parts of the trinity CFKB, including the Chinese function word usage dictionary, Chinese function word usage rule base, and Chinese function word usage corpus. In addition, we discussed the automatic identification of Chinese function word usages and the potential applications based on CFKB.

Next we will continue to improve the quality of CFKB, making sure the three parts of CFKB are in concordance. With these results, we will study the rule and statistic combined automatic identification algorithms of Chinese function word usages. And we will also attempt the applications based on CFKB, for example, adding the pre process or post process into the machine translation system to purify the quality of Chinese text. We hope CFKB would provide a solid data reference for Chinese information processing and its related applications.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 60970083), the Open Projects Program of National Laboratory of Pattern Recognition, and the Outstanding Young Talents Technology Innovation Foundation of Henan Province, China (No. 104100510026).

5 References

Burnard, Lou. 2000. Reference guide for the British National Corpus. Oxford University Computing Services, Oxford, UK.

Calvo, Hiram, Alexander Gelbukh, and Adam Kilgarriff. 2005. Distributional thesaurus versus WordNet: A comparison of backoff techniques for unsupervised PP attachment. In Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005). 177–188.

Chang, Pi-Chuan, Dan Jurafsky, and Christopher D. Manning. 2009. In Proceedings of the 47th Annual Meeting of the ACL(ACL 2009) and the 4th IJCNLP of the AFNLP. 215–223.

College Park, MD. Volk, Martin. 2003. German prepositions and their kin. A survey with respect to the resolution of PP attachment ambiguities. In Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications. 77 – 85.

Dong, Zhengdong, and Dong, Qiang. 1999. http://www.keenage.com/

Du, Jinhua and Andy Way. 2010. A Discriminative Latent Variable-Based "DE" Classifier for Chinese-English SMT. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). 286–294.

Fellbaum,C. 1999. WordNet : An Electronic Database, The MIT Press.

Gala, Nuria and Mathieu Lafourcade. 2005. Combining corpus-based pattern distributions with lexical signatures for PP attachment ambiguity resolution. In Proceedings of the 6th Symposium on Natural Language Processing (SNLP-05).

Gaussier, Eric and Nicola Cancedda. 2001. Probabilistic models for PP-attachment resolution and NP analysis. In Proceedings of the ACL/EACL-2001 Workshop on Computational Natural Language Learn (CoNLL-2001). 1 – 8.

Girju, Roxana. 2009. The syntax and semantics of prepositions in the task of automatic interpretation of nominal phrases and compounds: A cross-linguistic study. Computational Linguistics, 35(2):185 – 228.

Hao, Liping, et al. 2007. Research on Chinese adverb usage for machine recognition. In Proceedings of the 7th International Conference on Chinese Computing (ICCC 2007), 122-125.

Hartrumpf, Sven. 1999. Hybrid disambiguation of prepositional phrase attachment and interpretation. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99). 111 – 120,

Jørgensen, Fredrik and Jan Tore Lønning. 2009. A minimal recursion semantic analysis of locatives. Computational Linguistics, 35(2):229 – 270.

Kelleher, John D. and Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. Computational Linguistics, 35(2):271 – 306.

Kokkinakis, Dimitris. 2000. Supervised PP-attachment for Swedish: Combining unsupervised and supervised training data. Nordic Journal of Linguistics, 3(2):191 – 213.

Li, Xiaoqi, et al. 2005. The teaching materials of the modern Chinese function word. Peking University press, China. (in Chinese)

Litkowski, Kenneth C. and Orin Hargraves. 2007. SemEval-2007 task 06: Word-sense disambiguation of prepositions. In Proceedings of the 4th International Workshop on Semantic Evaluations, Prague. 24 – 29.

Liu, Rui, et al. 2008. The automatic recognition research on contemporary Chinese language, Computer Science, 8(A): 172-174. (In Chinese)

Liu, Yun. 2004. The Development of the Knowledge-base of Chinese Function Words. The postdoctoral report of Peking University.

Lu, Jianming, and Ma Zhen. 1999. Some comments on the modern Chinese function word. Chinese Press, China. (In Chinese)

Lv, Shuxiang. 1979. The problem on grammatical analysis of the Contemporary Chinese. Commercial Press, China. (In Chinese)

Lv, Shuxiang. 1980. Contemporary Chinese 800 words. Commercial Press, China. (in Chinese)

Miller, G.A. et al. 1993. Introduction to WordNet : An On-line Lexical Database. Specification of WordNet.

O' Hara, Tom and Janyce Wiebe. 2009. Exploiting semantic role resources for preposition disambiguation. Computational Linguistics, 35(2):151-184

Peng, Shuang. 2006. The Development of the Knowledge-base of Contemporary Chinese preposition. The postdoctoral report of Peking University.

Setiawan Hendra, Min-Yen Kan, Haizhou Li and Philip Resnik. 2009. Topological Ordering of Function Words in Hierarchical Phrase-based Translation. In Proceedings of the 47th Annual Meeting of the ACL(ACL 2009) and the 4th IJCNLP of the AFNLP. 324–332.

Timothy Baldwin, Valia Kordoni, and Aline Villavicencio. 2009. Prepositions in ppplications: a survey and introduction to the special. Computational Linguistics, 35(2):119-149.

van Herwijnen, Olga, Jacques Terken, Antal van den Bosch, and Erwin Marsi. 2003. Learning PP attachment for filtering prosodic phrasing. In Proceedings of the 10th Conference of the EACL (EACL 2003). 139 – 146.

Volk, Martin. 2006. How bad is the problem of PP-attachment? A comparison of English, German, and Swedish. In Proceedings of the Third ACL-SIGSEM Workshop on Prepositions. 81 – 88.

Wang, Hui, Weidong Zhan, and Shiwen Yu. 2003. The specification of semantic knowledge-base of contemporary Chinese. Journal of Chinese Language and Computing, 13 (2):159-176

Wang, Lei, and Shiwen Yu. 2010Semantic Computing and Language Knowledge Bases, In Proceedings of CIPS-SIGHAN Joint Conference on the 1st Chinese Language Processing (CLP 2010), 34-45

Yu, Jiangsheng and Shiwen Yu. 2002. The structure of Chinese concept dictionary. Journal of Chinese Information Processing. Vol.16 (4):12-20

Yu, Shiwen, et al. 1998. Introduction to Grammatical Knowledge Base of Contemporary Chinese. Tsinghua University Press, China. (in Chinese)

Yu, Shiwen, Xuefeng Zhu, and Yun Liu. 2003. Knowledge-base of generalized function words of modern Chinese. Journal of Chinese Language and Computing, 13(1):89-98.

Yuan, Yingcheng,, et al. 2010. The rule based algorithm design and its implementation of Chinese function word usages' automatic tagging. In Proceedings of the 11th Chinese Lexical Semantic Workshop (CLSW 2010). 163-169. (In Chinese)

Zan, Hongying, and Junhui Zhang. 2009. Studies on automatic recognition of Chinese adverb CAI's usages based on statistic. In Proceedings of the 3rd international conference on Natural Language Processing and Knowledge Engineering (NLPKE 2009). 393-397.

Zan, Hongying, Junhui Zhang, Xuefeng Zhu, and Shiwen Yu. 2010. Studies on automatic recognition of common Chinese adverbs usages based on statistics methods. In Proceedings of CIPS-SIGHAN Joint Conference on the 1st Chinese Language Processing (CLP 2010), 87-92

Zan, Hongying, Kunli Zhang, Yumei Chai, and Shiwen Yu. 2007. Studies on the function words knowledge base of modern Chinese. Journal of Chinese Information Processing, 21 (5):107-111. (in Chinese)

Zhang, Bin. 2001. Contemporary Chinese function words dictionary. Commercial Press, China. (in Chinese)

Zhou, Yihui, et al. 2010. An error driven method to improve rules for the recognition of Chinese modality "LE", In Proceedings of the 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLPKE 2010), 242-245.