ICIC Express Letters
Volume **10**, Number **6**, June **2016**

 ISSN 1881-803X
pp. **1511–1516**

AUTOMATIC EXTRACTION AND QUANTITATIVE ANALYSIS OF CHINESE SEPARABLE WORDS

Baorong He[1,2] and Likun Qiu[1,2,*]

[1]Key Laboratory of Language Resource Development and Application of Shandong Province
[2]School of Chinese Language and Literature
Ludong University
No. 186, Middle Hongqi Road, Zhifu District, Yantai 264025, P. R. China
[*]Corresponding author: qiulikun@gmail.com

Received November 2015; accepted February 2016

Abstract. *Chinese separable words is a research focus in Chinese linguistics and natural language processing. This paper proposed a method for extracting Chinese separable words both in their combined forms and separate forms. Based on the extracted results, we analyze unigram, bigram and trigram inserted elements and show that a multi-syllable inserted element usually composes two or more shorter ones. In addition, our analysis demonstrated that the separate degrees of separable words are quite different.*
Keywords: Separable word, Novel corpus, Inserted element, Separate ratio

1. **Introduction.** In Chinese, some verbs can be separated by inserting some types of elements. For instance, "*xizao* (take a bath)" and "*chikui* (take a beating)" can be transformed as "*xi ge zao* (take a bath) " and "*chi le kui* (took a beating)" by inserting "*ge* (a)" and "*le* (auxiliary word)", respectively [1]. After inserting some elements, the structures look like phrases, while they are typical words without inserting any elements. This type of verbs are defined as *separable words*. Separable words are characterized in that the two separate elements should be taken as a whole if one wants to understand the meaning of its separate form. That is, we cannot infer the meaning of a separate word by its two separate elements. For instance, the meaning of "*chikui* (take a beating)" are not equal to the combination of the meaning of the two elements "*chi* (eat)" and "*kui* (deficit)". To understand the meaning of a separate word, we should realize that the two separate elements are parts of a same word and then understand them as a whole. Because of this characteristics, the understanding and generation of separate words has been a research focus and difficulty in the domain of Chinese learning and natural language understanding.

For Chinese learners, the difficulty lies in exploring which words can be separated by inserting some elements, which types of inserted elements can be used for a certain word, and the construction meaning of a separate form. For natural language understanding, similarly, the difficulty lies in how to judge a structure is a separate form of a separate word and how to understand the structure. To address these problems, more research issues can be performed. In this paper, we only focus on analyzing the inserted elements and try to conclude the rules of separate words.

The rest of this paper is organized as follows. Section 2 describes the proposed method for extracting separate words both in separate form and combined form. Experimental results are showed in Section 3. Section 4 introduces related work. Finally, a brief conclusion is given in Section 5.

2. **Separable Word Extraction Method.** Given a raw corpus C and a bi-syllable word list W, the proposed extraction method is as follows.

First, a word segmentation and part-of-speech tagging tool is used to analyze C. With this tool, each sentence in C is segmented into a list of words, each of which is annotated with a part-of-speech such as verb (v), noun (n) and adjectives [2]. For instance, the analyzing result of "*laoshi* (teacher) *gen* (with) *Xiaowang* (Xiaowang) *tan le yi ci hua* (had a conversation)" is "*laoshi* (teacher)/n *gen* (with)/p *Xiaowang* (Xiaowang)/nr *tan*/v *le*/u *yi*/m *ci*/q *hua*/n (had a conversation)", where "n", "nr", "p", "v", "u", "m" and "q" indicate noun, personal name, preposition, verb, auxiliary word, numeral and classifier, respectively. In this sentence, the two elements of the word "*tanhua* (talk)" are separated by inserting a three-word inserted element "*le* (auxiliary word) *yi* (one) *ci* (time)". When we want to translate the separated structure, we have to take "*tanhua* (talk)" as a whole word to understand its actual meaning "talk".

Second, we extract separate form of each separable word from segmented and POS-tagged sentences. A two-level index I is built based on W. Given an analyzed sentence S and I, the extraction algorithm is shown in Figure 1. For each word $word_i$ in S, we judge whether its length is 1 and part-of-speech is verb (v), the part-of-speech of the left word $word_{i-1}$ is not verb, and $word_i$ is in the first-level of I. If all these constraints are satisfied, we then judge whether there is a word that contains one character, and is a verb, adjective or noun, and is in the second level of I (I[$word_i$]). If there exists one such word, a separable word $word = word_i + word_j$ and an inserted element $inserted\_elements = S$[i+1:j] are added into the separable word set *SWSet* and *IESet*, respectively. Using this algorithm, we process each sentence in corpus C, and acquire a large set of separable words and a corresponding set of inserted elements.

```
SWSet = {}, IESet = {}
for word_i in S:
    if LEN(word_i) = 1, POS(word_i) = "v", POS(word_{i-1})! = "v" and word_i in I:
        j = i + 2
        for word_j in S:
            if LEN(word_j) = 1, POS(word_j) = "a|v|n", and word_j in I[word_i]:
                word = word_i + word_j
                inserted_elements = S[i+1:j]
                i = j + 1
                AddToSet(SWSet, word)
                AddToSet(IESet, inserted_elements)
```

FIGURE 1. Separable word extraction algorithm

3. **Experiment.**

3.1. **Experimental setting.**

3.1.1. *Data.* Since the separate forms of Chinese separable words mainly appear in conversation, our analysis is based on contemporary Chinese novel. As for contemporary Chinese novel, we used several famous Chinese literary magazines such as *Fiction Monthly Magazine* and *Beijing Literature* published in 2012 to collect a corpus that contains 5.16 million words and 0.254 million sentences. We refer to the corpus as *Novel Corpus* [3].

As for candidate separable words, the most popular Chinese lexicon, i.e., *Contemporary Chinese Dictionary* (Fifth Edition) is used. In this lexicon, if a multi-syllable word is separable, the symbol "||" or "//" is inserted between the phonetic notations of the two parts of the word [4]. We collect a separable word list that contains 3937 separable words.

3.1.2. *Word segmentation and POS tagging.* We used the PMT version of ZPar, which can be downloaded from the Sourceforge website[1] [5,6], to perform word segmentation and POS tagging.

3.2. **Results and analysis.** The following analysis is mainly based on the *Novel Corpus*. Using the method described in Section 2, 13012 instances were extracted. In the following, we will analyze inserted elements and separate ratio, respectively.

3.2.1. *Inserted elements.* The number of the inserted element type appearing in all the instances is 4404. That is, each element appears in about 3 instances on average. The size distribution and frequency distribution of inserted elements are shown in Table 1 and Table 2, respectively. Here, "size" denotes the number of words in the inserted elements. For instance, "*le* (auxiliary word)", "*le* (auxiliary word) *shenme* (what)" and "*le* (auxiliary word) *yi kou* (a mouthful of)" are three inserted elements that contain one, two and three words, respectively. "Frequency" denotes the number of an inserted element appearing in the corpus.

Table 1 shows that the sizes of most inserted elements are less than 7, and the TOP 2 most popular sizes are 2 and 3. Similarly, from Table 2, we may find that the frequencies of most inserted elements are also less than 7. The two types of distributions differ in that the frequency distribution is a strict-decreasing distribution yet the size distribution is not.

Actually, a bigram inserted element (size = 2) is combined by two unigram inserted elements (size = 1), and a trigram inserted element (size = 3) is combined by one unigram inserted element and one bigram inserted element. Thus, we should analyze the three types of inserted elements, respectively.

Table 1. Size distribution of inserted elements

Size	1	2	3	4	5	6	7	> 7
Count	500	1073	1071	688	392	248	142	297

Table 2. Frequency distribution of inserted elements

Freq	1	2	3	4	5	6	7	$7 < \mathrm{F} \leq 20$	> 20
Count	3729	329	101	54	30	23	13	75	56

Unigram Inserted Elements. The frequency distribution of TOP 30 unigram inserted elements is shown in Table 3. Here, "IE" and "PL" denote inserted element and the part-of-speech list of an inserted element, respectively. "IF", "WTC" and "PLF" indicate the frequency of an inserted element, the count of the separable word types where an inserted element appears, and the frequency of a part-of-speech list, respectively.

As shown in Table 3, the TOP 30 inserted elements contain 6 auxiliary words (u), 5 classifiers (q), 6 verbs (v), 5 pronouns (r), 4 nouns (n), 2 numerals (m), 1 adverb (d) and 1 adjective (a). In addition, the WTCs of "*hua* (words)", "*zhi* (de)", "*dun* (a)", "*xin* (heart)", "*ju* (sentence)", *qi* (air) and "*lou* (floor)" are relatively less according to their IFs. This means that these inserted elements only co-occur with a small set of separable words. For instance, "*dun* (a)" only appears with the separable word "*chifan* (have a meal)" and thus forms the phrases "*chi dun fan* (have a meal)".

The TOP 20 bigram and trigram inserted elements are shown in Table 4 and Table 5, respectively. From Table 4, we may find that most bigram inserted elements are composed of two unigram inserted elements. For instance, "*le* (le) *kou* (a)" is composed of "*le* (le)" and "*kou* (a)", which are the first and twelfth unigram inserted elements, respectively.

[1]http://sourceforge.net/projects/zore/files/SegTagParsing/

TABLE 3. The frequency distribution of TOP 30 unigram inserted elements

IE	*IF*	*WTC*	*PL*	*PLF*	*IE*	*IF*	*WTC*	*PL*	*PLF*
le (le)	1276	447	u	671	*hua* (words)	56	1	n	84
zhe (zhe)	728	138	u	671	*dian* (a little)	52	35	q	158
bu (not)	697	58	d	75	*hao* (done)	47	30	a	73
guo (cross)	388	12	v	227	*xia* (down)	46	15	v	227
de (de)	385	103	u	671	*zhe* (this)	44	13	r	145
qi (begin)	365	47	v	227	*zhi* (de)	43	2	u	671
guo (guo)	345	134	u	671	*qi* (air)	34	3	n	84
ge (piece)	203	104	q	158	*zhexie* (these)	29	6	r	145
shenme (what)	150	67	r	145	*zhezhong* (this type)	28	10	r	145
wan (finish)	119	42	v	227	*dun* (meal)	27	1	q	158
de (de)	114	28	u	671	*zhege* (this)	26	18	r	145
kou (mouth)	97	8	q	158	*xin* (heart)	23	3	n	84
shang (get on)	80	28	v	227	*ju* (sentence)	21	4	q	158
chu (come out)	70	14	v	227	*yixia* (once)	19	14	m	67
yu (one)	57	24	m	67	*lou* (floor)	19	2	n	84

TABLE 4. The frequency distribution of TOP 20 bigram inserted elements

IE	*IF*	*WTC*	*PL*	*PLF*	*IE*	*IF*	*WTC*	*PL*	*PLF*
le (le) *kou* (a)	185	8	uq	67	*ni de* (your)	35	26	ru	74
le (le) *yi* (a)	101	16	um	81	*bu shang* (cannot)	35	16	dv	95
le (le) *ge* (q)	80	40	uq	67	*wo de* (my)	33	20	ru	74
le (le) *dian* (a little)	71	1	uv	30	*le* (le) *yixia* (once)	33	19	um	81
bu liao (unable to)	70	39	dv	95	*le* (le) *shenme* (what)	33	10	ur	58
bu (not) *chu* (out)	69	5	dv	95	*le* (le) *yige* (one)	31	10	um	81
guo (guo) *de* (de)	53	8	uu	21	*shang le* (done)	27	12	vu	93
bu (not) *guo* (guo)	48	5	dv	95	*ta de* (his)	25	18	ru	74
yi ju (a word of)	46	5	mq	113	*bu zhao* (cannot)	25	1	dv	95
yi kou (a mouthful of)	37	8	mq	113	*ji ju* (a few words)	25	1	mq	113

TABLE 5. The frequency distribution of TOP 20 trigram inserted elements

IE	*IF*	*WTC*	*PL*	*PLF*
le (le) *yi kou* (a mouthful of)	132	8	umq	118
le (le) *yi ju* (a word of)	26	4	umq	118
le (le) *wo de* (my)	15	6	uru	37
wan (finish) *zhe ju* (this)	15	2	vrq	8
le (le) *ta de* (his)	14	9	uru	37
chu (out) *yi kou* (a mouthful of)	13	3	vmq	24
guo (guo) *yi ju* (a word of)	13	2	umq	118
zhe (zhe) *ta de* (his)	12	6	uru	37
zhe (zhe) *ta de* (her)	12	5	uru	37
le (le) *ji ju* (several words of)	12	1	umq	118
guo (guo) *yi ci* (one time)	11	11	umq	118
le (le) *ji ci* (several times)	11	11	umq	118
de (de) *yi ju* (a sentence)	11	1	umq	118
de (de) *na ju* (that)	9	1	urq	20
le (le) *ji ge* (several)	8	6	umq	118
chu (out) *zhe ju* (this)	7	3	vrq	8
qi (begin) *zhe jian* (this)	7	1	vrq	8
le (le) *ta de* (her)	6	6	uru	37
le (le) *yi ci* (one time)	6	4	umq	118
zhe (zhe) *ziji de* (own)	6	3	uru	37

Further, Table 5 demonstrates that most trigram inserted elements are composed of a high-frequency unigram inserted element and a high-frequency bigram inserted elements. For instance, "*le* (le) *yi kou* (a mouthful of)" is composed of "*le* (le)" (a unigram inserted element) and "*yi kou* (a mouthful of)" (a bigram inserted elements).

3.2.2. *Separate ratio.* Some typical separable words tend to appear in natural text in their separate forms, while some other separable words appear in their combined forms in most cases. We use separate ratio (SR) to measure the typicality of a separable word. SR is computed according to Equation (1), where SC and CC indicate the count of a separate word in its separate form and combined form, respectively. We list the TOP 20 high-frequency separable words in Table 6. This table shows that the separate degrees of different separable words vary considerably. Some words such as "*shuoshi* (focus)", "*songqi* (relax one's efforts)" and "*biyan* (close one's eyes)" mainly appear in their separate form; some other words such as "*kanjian* (see)" and "*huiqu* (return)" mainly appear in their combined form; while both two forms of some words are similarly popular.

$$SR = \frac{SC}{SC + CC} \tag{1}$$

Table 6. Separate ratios of TOP 20 high-frequency separable words

Word	*SC*	*CC*	*SR*	*Word*	*SC*	*CC*	*SR*
shuohua (say)	1042	1991	0.34	*huitou* (turn round)	137	336	0.28
shuoshi (focus)	343	0	1	*tinghua* (tractable)	137	98	0.58
tanqi (sigh)	308	47	0.86	*zoulu* (walk)	131	180	0.42
chifan (have a meal)	287	912	0.23	*chushi* (have an accident)	115	201	0.36
taitou (raise one's head)	273	374	0.42	*diantou* (nod)	114	592	0.16
zuoshi (handle affairs)	268	146	0.64	*shuijiao* (sleep)	104	498	0.17
kanjian (see)	249	2207	0.10	*huiqu* (return)	102	1272	0.07
chulai (come out)	175	5493	0.03	*songqi* (relax one's efforts)	101	1	0.99
zhuanshen (turn-back)	146	460	0.24	*kaiche* (drive)	99	348	0.22
ditou (lower one's head)	146	273	0.34	*biyan* (close one's eyes)	93	0	1

3.3. **Evaluation.** We evaluated all the 13012 instances manually. The overall precision is 75.9%. If we only evaluate the instances that occur more than once, the precision is 83.8%. Further analysis shows that the precision of the instances that contain more popular inserted elements (WTC $>$ 2) is much higher than that of the instances that contain less popular inserted elements (WTC $<=$ 2). In addition, when the size of inserted elements is larger than 5, the inserted elements that contain the character "*de* (of)" usually mean higher precision than those without the character.

4. **Related Work.** Wang [1] discussed the separate causes for Chinese separable words. He believed that it is mainly because of that phrases and words in Chinese share similar structure rules. Thus, if we want to express information of aspect and amount of action, our only choice is to insert some elements into a word, just like inserting aspect markers such as "*zhe* (zhe)", "*le* (le)" and "*guo* (guo)" into a verb-object phrase such as "*kan* (see) *dianying* (a movie)".

The research of Ren and Wang [7] is corpus-based. They collected 423 separable words, analyzed three types of inserted elements, and found 12 typical separable words, including "*tinghua* (tractable), *fenqing* (distinguish), *chushi* (have an accident), *xiyan* (smoke), *shangdang* (be tricked), *ganbei* (cheers), *chijing* (amaze), *dazhang* (fight a battle), *chikui* (suffer losses), *fanzui* (commit a crime), *woshou* (shake hands), *xiangfu* (enjoy a happy life)". This separable word list is quite different from ours (Table 6), because our original

word lists and corpora are different. Our work also differs in that we focus on analyzing the relation between longer inserted elements and shorter elements.

Wang [8] focused on investigating the distribution difference of separate forms of Chinese separable words in different genres. They concluded that separate forms often appear in colloquial texts such as novel and drama, and the degrees of formality and subjectivity are two main factors that determine whether separate forms can be used.

Shi [9] classified Chinese separable verbs into four categories, including *verb+noun*, *verb+complement*, *verb+verb* and *verb+complement+noun*. He further discussed the analysis, representation and translation strategies for different types of separable verbs.

5. **Conclusions.** This paper proposed a method for extracting Chinese separable words both in their combined forms and separate forms. Based on the extracted results, we analyzed unigram, bigram and trigram inserted elements and show that a multi-syllable inserted element is usually composed of two or more shorter ones. In addition, we show that the separate degrees of separable words are quite different.

We also observed that different types of separable words are related with different types of inserted elements. For instance, the inserted element "*bu* (not)" usually appears in verb-complement verbs such as "*kanjian* (see)" and "*chulai* (come out)" yet cannot appear in verb-object verbs such as "*chifan* (have a meal)" and "*taitou* (raise one's head)". We will discuss this issue in future work.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (No. 61572245 and No. 61103089), Scientific Research Foundation of Shandong Province Outstanding Young Scientist Award (No. BS2013DX020), and Humanities and Social Science Projects of Ludong University (No. WY2013003).

REFERENCES

[1] H. Wang, Discussion on separate causes for Chinese separable words, *Linguistic Researches*, no.3, pp.29-34, 2002.

[2] S. Yu, H. Duan, X. Zhu, B. Swen and B. Chang, Specification for corpus processing at Peking university: Word segmentation, pos tagging and phonetic notation, *Journal of Chinese Language and Computing*, vol.13, no.2, pp.121-158, 2003.

[3] L. Qiu and S. Kang, Construction of a Chinese balanced corpus for language teaching and dictionary compilation, *Journal of Modernization of Chinese Language Education*, vol.3, no.1, pp.31-36, 2014.

[4] Dictionary Editorial Office, Institute of Linguistics, Chinese Academy of Social Science, *Contemporary Chinese Dictionary*, 5th Edition, Commercial Press, 2005.

[5] Y. Zhang and S. Clark, Syntactic processing using the generalized perceptron and beam search, *Computational Linguistics*, vol.37, no.1, pp.105-151, 2011.

[6] L. Qiu and Y. Zhang, ZORE: A syntax-based system for Chinese open relation extraction, *Proc. of EMNLP*, Doha, Qatar, pp.1870-1880, 2014.

[7] H. Ren and G. Wang, Corpus analysis on Chinese separable words, *Linguistic Science*, vol.4, no.6, pp.75-87, 2005.

[8] H. Wang, Study on stylistic differences of Chinese separable words, *Applied linguistics*, no.3, pp.81-89, 2009.

[9] X. Shi, Treatment of separable words in Chinese-English machine translation, *Proc. of Machine Translation Workshop*, pp.68-76, 2002.