WORD SKETCHES OF SEPARABLE WORDS *LIHECI* IN CHINESE

Mateja PETROVČIČ
University of Ljubljana, Faculty of Arts
mateja.petrovcic@ff.uni-lj.si

Abstract

Separable words (*liheci*) are a special type of Chinese verbs with unique syntactical features in a sense that some elements come in between the two morphemes of a verb for a sentence to be grammatically acceptable. Not all separable words are extendable to the same degree. To understand the behaviour of words, it is generally advised to check word sketches, because they are based on large text corpora. This article examines how Chinese separable words are treated in Sketch Engine and discusses on the appropriateness of the available Chinese corpora for word sketches. It further stresses the importance of including information on inserted elements in word sketches and gives suggestions on how to include them.

Keywords: Chinese; separable words; *liheci*; word sketches; Sketch Engine

Povzetek

Ločljive besede (*liheci*) so poseben tip glagolov s svojevrstnimi sintaktičnimi lastnostmi. Zanje je značilno, da moramo določene stavčne člene vstaviti med oba morfema glagola. Pri ločljivih besedah je posebej problematično to, da ne moremo nikoli natančno vedeti, do kakšne mere je beseda ločljiva in s katerimi vzroci jo lahko razširimo. Pri orisu rabe besed so nam lahko v veliko pomoč besedne skice, ki izhajajo iz besedilnih korpusov. V članku proučimo, kako ločljive besede obravnava Sketch Engine, kateri kitajski korpus je najbolj primeren za besedne skice in predlagamo, kako bi besednim skicam dodali tudi informacije o ločljivosti glagolov.

Ključne besede: kitajščina; ločljive besede; *liheci*; besedne skice; Sketch Engine

1 Introduction

Separable words in Chinese have been investigated by numerous researchers for several decades. According to their research orientation and focus, Huang (2006) divides them into two periods. The first period lasted from 1950s to 1970s, and the second period stretches from the 1980s to the present.

Huang (2006) notes that due to the problems related to transliteration this group of verbs caught linguists' attention even before the term *separable word* (*liheci* 离合)

Acta Linguistica Asiatica, 6(1), 2016.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.6.1.47-57

was defined. Chinese writing system does not have explicit word boundary markers, such as the spaces between words. When the official romanization system for Standard Chinese *Hanyu pinyin* was to set orthographic rules, it encountered the problem of word boundaries. Even after several decades of discussions, scholars had yet to agree whether these "items" are words, word phrases, words as well word phrases, or something in-between.

After the China's opening up policy in 1980s, learning and teaching Chinese as a foreign language became the major point of interest. Separable words became the subject of research in relation to foreign language acquisition. They seemed to be a common problem for foreign students regardless of their native language. It was obvious that a non-native speaker had difficulties understanding whether a verb should be used as a unit or separately, and if the latter, which elements could be inserted in between the two morphemes (Huang, 2006).

This second period has also been related to the development of information processing and machine translation, which brought new insights into the existing research topics. Analyzing language by means of corpus linguistics is also one of the novelties in this period. Several recent papers refer to the data from Peking University CCL Online Corpus.

2 Separable words *liheci*

Separable words are disyllabic verbs that are separable in certain circumstances. Even more, in these circumstances, some separable words should undertake at least one element in between its syllables (morphemes), or else the sentence would be grammatically incorrect. There are several types of separable verbs, but the majority of them has the morphological structure "verb–object", for example *tiao//wu* (跳舞) "to dance" (lit. to jump dance), *jian//mian* (见面) "to meet" (lit. to see a face), *bang//mang* (帮忙) "to help" (lit. to help//busy).

4. 他 **跳** 了 一个小时 的 **舞**。
   ta **tiao** le yi ge xiaoshi de **wu**
   he **dance** LE one hour DE **dance**
   He was dancing for an hour.

5. 我们 只 **见** 过 一次 **面**。
   women zhi **jian** guo yi ci **mian**
   we just **see** GUO once **face**
   We've met only once.

6. 他 **帮** 了 我 一个 大 **忙**。
   ta **bang** le wo yi ge da **mang**
   he **help** LE I one big **help**
   He helped me a lot.

Scholars advocate separable words in two ways; either as *words* or as *word phrases*. Those who interpret them as *words* support their ideas with the following three facts. Firstly, separable words may be uttered in isolation with semantic or pragmatic content; secondly, several morphemes of separable words are bound morphemes; and finally, although separable words can be extended, their extension patterns are very limited. On the other hand, scholars who claim that separable words are *word phrases* say that separable words carry syntactic features of word phrases such as splitting, flexible word order, and often carry a special idiomatic meaning (Zhou, 2010, p. 123).

Different authors propose various categorizations of separable words, classifying them into up to ten different groups. However, it is generally agreed that there are at least the following three types of separable words (Huang, 2006, p. 85):

- V–O type (*dongbin shi* 动宾式)
- V–Complement type (*dongbu shi* 动补式)
- S–V type (*zhuwei shi* 主谓式)

As mentioned above and demonstrated in examples 1–3, the two morphemes of a separable word may demand one or more additional elements in between them. Such additional elements may be an aspectual particle and a durational phrase, as in example 1, an aspectual particle and a phrase expressing number of occurrences, as in example 2, or others. Example 3 is extended with an aspectual marker, followed by the recipient of an action and an attribute.

Scholars have come to several conclusions on which elements can be inserted in separable words. In a very simplified manner, we present Zhou's (2010) conclusions because they are well organized and systematic.

- aspectual particles (*le* 了, *guo* 过 and *zhe* 着)
- complements (quantitative, resultative, directional, potential)
- attributes
- some question forms and patterns
- a combination of these elements

The most intriguing part concerning the inserted elements is their degree of separability. Some separable words can be extended with all the above patterns whereas others are limited to some of them (He, 2009, p. 65). Wang (2008; 2010) provides us with corpus–driven findings, where he concludes that the majority of separable words are related to our everyday's life and activities. Such separable words are also very flexible and allow various combinations of extension. Wang draws insightful conclusions about the semantics of separable words but due to length limitations of this paper, we will not go into details.

In this paper, we will rather focus on word sketches, which may provide collocational and grammatical features of words.

3 Word sketches

Following the explanation on the Sketch Engine's website, a word sketch is "a corpus–based summary of a word's grammatical and collocational behavior" (Getting Started with Sketch Engine, 2016). This is a very useful tool not only for researchers, but also for language teachers, language learners and other users, because it shows "the word's collocates categorized by grammatical relations such as words that serve as an object of the verb, words that serve as a subject of the verb, words that that modify the word etc." (Word Sketch, 2016).

Sketch Engine may include several corpora for the same language. For standard Chinese, there are nine text corpora available for subscribed users.[1] However, their word sketches vary remarkably. The main reason for such divergence is not the size of the underlying corpora but the availability of syntactical descriptions. Namely, word sketches depends on the available grammatical definitions supplied to Sketch Engine (Getting Started with Sketch Engine, 2016).

Figure 1 shows three sets of grammatical relations that are available for Chinese corpora. From the user's perspective, these can be selected from the *advanced options* of word sketch tool.

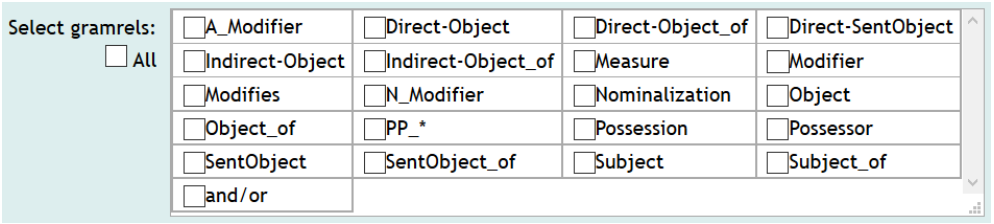

Figure 1a: Chinese grammatical relations (Set 1)

Figure 1b: Chinese grammatical relations (Set 2)

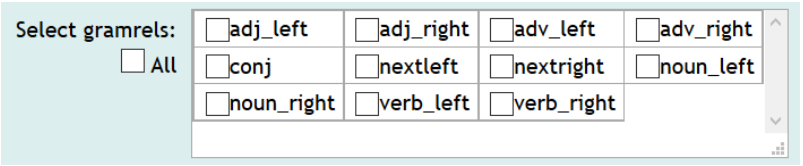

Figure 1c: Chinese grammatical relations (Set 3)

[1] The list of Sketch Engine's text corpora is available at https://www.sketchengine.co.uk/corpora/.

Grammatical relations are defined as regular expressions over Part-of-speech-tags (POS-tags), and are saved in the so-called *gramrel files*.[2] They are typically created for nouns, verbs, and adjectives, but can be enriched with other definitions, as well.

Presently, the best Chinese gramrel file is related to the Chinese GigaWord 2 corpus, both the Mainland (simplified) version and the Taiwan (traditional) version. Corpus zhTenTen [2011] is compared to other Chinese corpora in Sketch Engine much larger, and would therefore generate better word sketches, but has less sophisticated definitions for grammatical relations. Its wordsketches are therefore not as informative as in Chinese GigaWord 2 corpus (see Table 1).

Table 1: Chinese corpora and their corresponding grammatical relations

Text corpus	**Number of tokens**	**Grammatical relations**
Chinese GigaWord 2 Corpus: Mainland, simplified	299,338,099	Set 1, Figure 1a above
Chinese GigaWord 2 Corpus: Taiwan, traditional	455,526,209	Set 1, Figure 1a above
zhTenTen [2011]	2,106,661,021	Set 2, Figure 1b above
OPUS2 Chinese Simplified	299,338,099	Set 2, Figure 1b above
OPUS2 Chinese Traditional	622,382	Set 3, Figure 1c above
Internet-ZH	277,931,664	N/A
ChineseTaiwanWaC	349,198,060	Set 2, Figure 1b above
ChineseTaiwanWaC (Universal Sketch Grammar)	349,198,060	Set 3, Figure 1c above

[2] Gramrel file related to Figure 1a:
https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/cgw2_sc

Gramrel file related to Figure 1b:
https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/opus2_zh_TW

Gramrel file related to Figure 1c:
https://the.sketchengine.co.uk/bonito/corpus/wsdef?corpname=preloaded/opus2_zh

4 Word sketches of separable words

Although word sketches are primarily created for nouns, adjectives and verbs, information on grammatical and collocational behavior of separable words is still very limited. Recall that separable words are a special type of verbs.

In this research, we focus on Chinese GigaWord 2 Corpus, because it provides the most comprehensive word sketches among Chinese corpora. Queries have shown that separable words are treated as words in their disyllabic form, but have been assigned different POS-tags. Table 2 presents categorization of 21 basic separable words, listed on HSK3 vocabulary list.[3]

Table 2: Separable words and their POS-tags (HSK3)

Verb	Pinyin	Literal meaning[4] → Meaning	POS-tag[5]
睡觉	shui//jiao	to sleep//a sleep → to sleep	VA12
刮风	gua//feng	to blow//wind → to blow	VA3
见面	jian//mian	to see//a face → to meet	VA4
结婚	jie//hun	to tie//a marriage → to marry	VA4
跑步	pao//bu	to run//a step → to run	VA4
起床	qi//chuang	to get up//a bed → to get up	VA4
上网	shang//wang	to go up//a net → to go online	VA4
说话	shuo//hua	to speak//words → to speak	VA4
跳舞	tiao//wu	to jump//a dance → to dance	VA4
洗澡	xi//zao	to bathe//a bath → to bathe	VA4
游泳	you//yong	to swim//swimming → to swim	VA4, VA
帮忙	bang//mang	to help//busy → to help	VC2
离开	li//kai	to depart//to start → to leave	VC2
完成	wan//cheng	to finish//to complete → to complete	VC3
生病	sheng//bing	to arise //illness → to fall ill	VH11

[3] HSK stands for "Hanyu Shuiping Kaoshi" or Chinese Proficiency Test. HSK3 is the third of six levels.

[4] Literal meaning of each morpheme is provided for better understanding of the corresponding bisyllabic word.

[5] List of POS-tags for Chinese: https://www.sketchengine.co.uk/symbols-of-parts-of-speech/

Verb	Pinyin	Literal meaning[4] → Meaning	POS-tag[5]
发烧	fa//shao	to dispatch//heat → to have a fever	VH11
生气	sheng//qi	to arise//steam → to be angry	VH21
着急	zhao//ji	to take action//urgent→ to worry	VH21
担心	dan//xin	to carry//a heart → to be anxious	VK1
放心	fang//xin	to put down//a heart → to be at ease	VK1
注意	zhu//yi	to focus // an idea → to pay attention	VK1

Although the list is very short and as such not the best representative sample of separable words in standard Chinese, we can see at a glance that the major part of separable words is tagged as VA4. Based on this idea, I have further analyzed Wang's (2008) list of 207 separable words and got roughly similar results, as shown in Table 3.

Table 3: POS-tagging for 207 separable words

POS-tag	Number of separable words	Percentage
/VA4/	80	39%
/VH11/	47	23%
/VB11/	10	5%
/VH21/	9	4%
/VB12/	7	3%
/VA13/	6	3%
/VK1/	5	2%
/VC2/	4	2%
/VC31/	4	2%
Sub Total	**172**	**83%**
Others (less than 1 % each)	35	17%
Total	**207**	**100%**

In overall, the results show that almost 62% of all separable words are classified as VA4 or VH11. We assume that this percentage is even higher if we eliminate some "suspicious" items. For example, it is highly disputable whether 注意 "to pay attention"

is a separable word or not. The dictionary of 5000 graded words for New HSK provides an example of separate usage, but this is not very common.

你身体不好，健康状态要多**注**点儿**意**。
Ni shenti bu hao, jiankang zhuangkuang yao duo **zhu** dianr **yi**.
You are in poor health. Take care of your health condition. (Li, 2013, p. 381)

However, most of other authors do not consider this verb as separable. No cases of separable use were found in Chinese GigaWord 2 Corpus, nor in CCL corpus (Wang, 2008). Furthermore, even if this verb can be used separately, such examples are probably not frequent enough to be relevant for word sketches.

We have already noted (see Chapter 2 above) that in certain patterns, some elements must be inserted between the first and the second morpheme. Although this is a very important syntactical feature of Chinese separable words, word sketches offer no such information.

Based on corpus query language (CQL), we further analyzed the selected 21 separable words in their separate forms. We formulated the following CQL expression:

"*A*"[word!="\，|\;|\：|\。|\？|\！|\)" & tag!="PARENTHESISCATEGORY"]{1,}"*B*" within <p/>[6]

Query results were very fruitful and relatively accurate. The concordance list included all desired extensions, mostly without noise. Figure 1 shows one segment of the results for verb *bangmang* 帮忙 "to help".

	word	Frekvence
p | N	帮 了 大 忙	67
p | N	帮 个 忙	16
p | N	帮 大 忙	12
p | N	帮 了 我们 大 忙	12
p | N	帮 了 忙	12
p | N	帮 了 我 的 大 忙	8
p | N	帮 的 忙	7
p | N	帮 这 个 忙	5
p | N	帮 了 我们 的 大 忙	5
p | N	帮 了 他 的 忙	5
p | N	帮 了 他 一 个 大 忙	4
p | N	帮 点 忙	3
p | N	帮 什么 忙	3
p | N	帮 了 我 的 忙	3
p | N	帮 了 我 一 个 大 忙	3
p | N	帮 了 很 大 的 忙	3
p | N	帮 了 他 的 大 忙	3
p | N	帮 了 他 一点 忙	3
p | N	帮 过 忙	2
p | N	帮 俺 忙	2

Figure 1: Collocations of verb *bangmang*'s extended form

[6] Letters *A* and *B* represent the *first* and *second* morpheme of a separable word in question.

Among other results it is worth mentioning an example with 11 tokens inserted between both parts of the separable word, as shown and explained in Figure 2. Despite a far distance between the two morphemes, the structure shows syntactically correct relation.

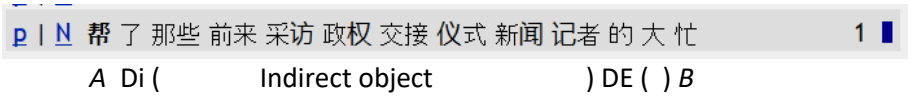

Figure 2: Collocations of verb *bangmang*'s extended form

However, mMorphemes of separable words do not tend to be so far apart as in the example from Figure 2. Analysis has shown that there are usually one to five tokens in between (Figure 3).

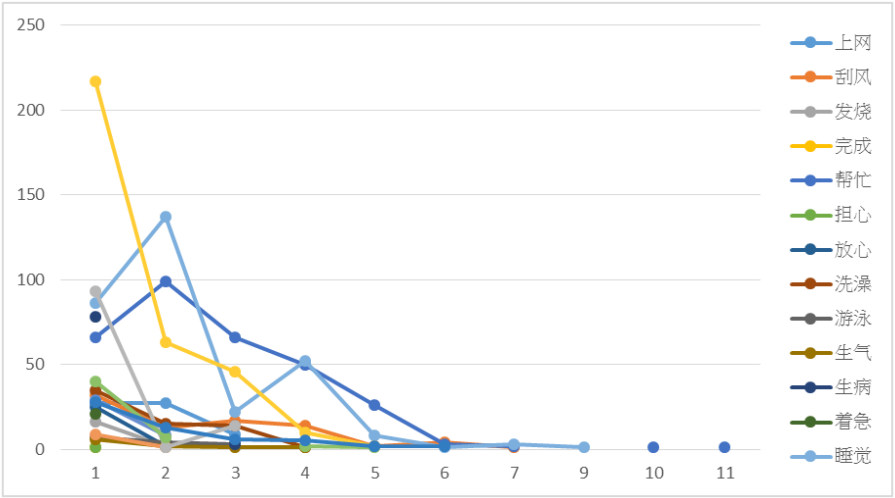

Figure 3: Number of inserted tokens in separable words

Separable words in their separate forms are generally treated individually and are not considered as the same verb anymore. Therefore We further investigated which part-of-speech tags are assigned to such separate forms and the results are shown in Table 4. Because our focus is not on the meaning of every single separable word but rather on their associated POS-tags, we intentionally left out English expressions.

Table 4: POS-tags of separable forms for HSK3 separable words

SW as a unit	SW in its separate form
睡觉/VA12	睡/VA12 [...] 觉/Nad; /VK1
刮风/VA3	刮/VC2 [...] 风/Na

SW as a unit	SW in its separate form
见面/VA4	见/VE2 […] 面/Na; /Ncda; /Nfa
结婚/VA4	结/VC31 […] 婚/Nad
跑步/VA4	跑/VA11 […] 步/Nf
起床/VA4	起/VC31; /Di; / Ng […] 床/Nab; /Nfa
上网/VA4	上/VC1; /Ng; /Nes […] 网/Nab
说话/VA4	说/VE2 […] 话/Nac
跳舞/VA4	跳/VA11; /Na […] 舞/Nac; /VC2
洗澡/VA4	洗/VC2 […] 澡/Na
游泳/VA4; /VA	游/VA11; /Nbc […] 泳/b
帮忙/VC2	帮/VC2; /P37 […] 忙/VH11
离开/VC2	离/VC2 […] 开/VC
完成/VC3	完/VH11 […] 成/VH11
生病/VH11	生/VC31 […] 病/VH11
发烧/VH11	发/VH11; /VD1; /VJ […] 烧/VC2
生气/VH21	生/VC31 […] 气/Naa; /VK
着急/VH21	着/VC2 […] 急/VH
担心/VK1	担/VC2 […] 心/Na
放心/VK1	放/VC33 […] 心/Na
注意/VK1	N/A

In 6 cases, the first morpheme is interpreted as a VC2 verb, in 4 cases as a VC31 verb, and in 3 cases as a VA11 verb. There are further 2 instances of a VE2 and VH11 verb, and the remaining 4 verbs merged to some minor groups.

The second morpheme is a noun in most cases, as expected. Recall that the majority of separable words have the internal morphological structure verb–object (V–O). The HSK3 list of separable words is too short to draw reliable conclusions, but we do believe that the above findings show correlalations and patterns that could, with some further support, get generalized.

5 Final thoughts

The preliminary research on separable words in Chinese GigaWord 2 corpus has shown that separable words are not treated as a special subtype of verbs. Instead, the disyllabic verb and its monosyllabic counterpart are tagged as two different verbs,

usually belonging to different POS-tags. It is impossible to expect POS-tags to be changed, however, it would be possible to track the relations between disyllabic and monosyllabic couterparts. To do so, a large number of separable words should be analyzed.

Collocations are therefore generated separately for disyllabic and monosyllabic forms of these verbs. Consequently, word sketches do not provide information about which elements should be inserted between the two morphems of separable words. This is undoubtly a very important syntactical feature of Chinese verbs.

Since CQL queries provide quite accurate results, and it is already known which patterns may be formed with separable words, it might be possible to create additional definitions of grammatical relations for the relevant gremrel file.

References

Getting Started with Sketch Engine. (2016, 3 12). Retrieved from Sketch Engine: https://www.sketchengine.co.uk/getting-started/

He, Q. [何清强]. (2009). Influence of Separation Extent on Acquisition of Verb-object Separable Words [分离度对动宾式离合词习得的影响]. *Journal of Ningbo University (Liberal Arts Edition), 22(6)*, pp. 65–70.

Huang, X. [黄晓琴]. (2006). A Summery of the Research on Separable Word [离合词研究综述]. *Journal of ILi Normal University*, pp. 84–87.

Li, L. [李禄兴]. (2013). *A dictionary of 5000 graded words for new HSK (Levels 1, 2 & 3)* [新 HSK 5000 词分级词典（一~三级）]. Beijing: Beijing Language and Culture University Press.

Wang, H. [王海峰]. (2008). *The Study on the Separable Words Separated Form Function of Mandarin Chinese* [现代汉语离合词离折形式功能研究]. *PhD thesis.* Beijing: Beijing Yuyan Daxue.

Wang, H. [王海峰]. (2010). A corpus-based semantic study of the Chinese separable words [基于语料库的现代汉语离合词语义特征考察]. *Journal of Hebei Normal University (Philosophy and Social Sciences Edition), 33(1)*, pp. 96–100.

Word Sketch. (2016). Retrieved from Sketch Engine: https://www.sketchengine.co.uk/word-sketch/

Zhou, W. [周卫华]. (2010). Expanding Forms and Features of Separable Words in Modern Chinese [现代汉语离合词的扩展形式及特点]. *Sanxia luntan, 6*, pp. 123–127.