Research report of
A corpus-based study of split words in Chinese
(RES-000-22-2286)

1. Background

According to Shi (1999, p. 123), there are 2,960 Splittable Compounds (SCs hereafter, better known among Chinese linguists as Lihe Ci, ‘Separate-combined Words’) entries in the Modern Chinese Dictionary (1996 edition), which account for 51.63% of all verb object words in the dictionary. SCs in Chinese morphology are hardly ever an undisputed issue in view of their unclear morpho-syntactic status.

Chinese SCs consist of two parts, the Head (SC_h) and the Tail (SC_t), which can be used as separate lexemes but have a distinct semantics when used together in a compound. In Chinese, however, the head and tail of the SC can be juxtaposed to form one composite word, as in (1) or separated by some intervening grammatical elements, as in (2) while maintaining the distinct semantics of the compound.

(1) dan1xin1, carry heart, ‘to worry’
(2) dan1-le yi1shang4wu3 xin1, carry ASP one morning heart, ‘to be worried the whole morning’

Thus whereas dan1 and xin1 when used on their own as separate lexemes mean “to carry” (as in dan1shui3, ‘carry water’, dan1chai2, ‘carry firewood’, etc.) and “heart” respectively, in both (1) and (2) the meaning “worry” is preserved. In fact, in some cases, the meaning “to worry” still obtains even if xin1 precedes rather than follows dan1 within the boundary of the same clause or sentence, as in (3).

(3) xin1yi4zhi2 dan1-zhe, heart all the time carry ASP, ‘to have been worried ever since’

The two morphological elements dan1 and xin1 actually glue together around the inherent semantic content constituted by the two units.

Chinese grammarians have been in disagreement with each other for over half a century as to whether SCs are words or phrases, or, following Lu (1957) who coined the term Lihe Ci, something in between (cf. Chao, 1968; Hu and Fan, 1996; Li and Thompson, 1981; Lü, 1979; etc). We have sought to resolve some of the controversies surrounding SCs in Chinese by providing a detailed corpus-based investigation of their distribution, structure and functions.

2. Research objectives
We have conducted both quantitative and qualitative analyses to address the following research questions regarding the linguistic behaviours of SCs:

1) to examine the types of interposing elements and their syntactic functions;
2) to identify the typical contexts of combined and separated uses of SCs;
3) to examine the interaction of SCs with syntactic features;
4) to explore the discourse-pragmatic functions of SCs;
5) to measure the “splittability” of the most commonly used SCs;
6) to examine genre variation in terms of the use of SCs;
7) to achieve a better understanding of the typology of morpho-syntactic types;
8) to explore whether the same phenomenon occurs in other isolating languages, be it in Southeast Asia or other areas.

The major objectives have been achieved in our forthcoming article to appear in Language Sciences. The paper describes in detail the morpho-syntactic contexts and the discourse-pragmatic functions of SCs. The splittability of SCs was studied by looking at the split use of SCs and their non-split versions in the same datasets. In the article, we find that both the internal and external structure analysis of discontinuous SCs suggests that the split SC usage is not sensitive to its sentential context in general. Rather, the structures and functions of the insertions within SCs have more weight in keeping verbal heads and nominal components apart. SCs are found to be significantly more frequently used in spoken, narrative, and involved texts. This genre preference is particularly characteristic of the split use of SCs. At two international conferences (Creating Infrastructure for Canonical Typology in Surrey, England, and Verb Typologies revisited: A Cross-linguistic Reflection on Verbs and Verb Classes in Ghent, Belgium), we attempted to extend our corpus method to discontinuous morpho-syntactic constructions in other languages. Due to the limited resources of Southeast Asian languages, we have not explored discontinuous verbal constructions with similar linguistic properties. Our analysis of SCs informed our compilation of Routledge Frequency Dictionary of Mandarin Chinese, and it is particularly helpful to pinpoint the disyllabic verbs. This research and the dictionary will certainly be a refreshing resource for learners of Mandarin.

3. Data and method

We chose the corpus-based method to investigate SCs, because corpus data and the related methodology have been increasingly recognised as an important resource in linguistic research and are used in nearly all branches of linguistics.

3.1 The corpora

Two corpora have been used in this project: the Lancaster Los Angeles Corpus of Spoken Chinese (LLSCC) for spoken Chinese and the Lancaster Corpus of Mandarin Chinese (LCMC) for written Chinese. The two million words datasets formed the empirical basis for our quantitative and qualitative analysis of SCs.

3.2 The procedures
First of all, we used the 1,738 commonly used splittable compounds listed in A Dictionary of Splittable Compound Usage in Modern Chinese (Yang, 1995) as seeds to automatically extract all instances of SCs when their two parts are separated, in either forward or backward direction, by 1-10 tokens. The results of our corpus search confirm that only 166 SC types out of the 1,738 SCs were found in the two million words of data.

A total of 2,793 concordance lines of SCs were returned from the 2 million word corpora with a Perl script that we wrote for this project, among which 1,348 instances are crude SCs in the written corpus of the LCMC, and 1,445 instances of crude SCs in the spoken corpus of the LLSCC.

Next we manually removed noise cases. After manual filtering, only 565 legitimate SCs were left for further morpho-syntactic and pragmatic annotation and analysis. The distribution of the true SCs in the two corpora is shown in Table 1.

<table>
<thead>
<tr>
<th>Corpora</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLSCC</td>
<td>108</td>
<td>327</td>
</tr>
<tr>
<td>LCMC</td>
<td>104</td>
<td>238</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>166</strong></td>
<td><strong>565</strong></td>
</tr>
</tbody>
</table>

Table 1. Distribution of SCs in LLSCC and LCMC

Then the concordances were annotated in terms of insertion type, the direction of separation, semantic type, pragmatic meaning, clause type, discourse function, and genre, etc. All the annotation was double checked to ensure inter-annotator reliability. Once the data had been thus annotated, we carried out the quantitative and qualitative analyses (see the Appendix I for a full explanation of the annotation scheme).

4. Results, findings and discussions

4.1 General distribution

Table 2 demonstrates the overall occurrences of SCs in the spoken and written corpora.

<table>
<thead>
<tr>
<th></th>
<th>LLSCC 1,002,151</th>
<th>LCMC 1,006,731</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC types</td>
<td>108</td>
<td>104</td>
<td>0.09</td>
</tr>
<tr>
<td>SC tokens</td>
<td>327</td>
<td>238</td>
<td>14.49</td>
</tr>
</tbody>
</table>

Table 2. Overall frequencies of SC types and tokens in LLSCC and LCMC

<table>
<thead>
<tr>
<th>Freq.</th>
<th>SCs</th>
<th>Freq.</th>
<th>SCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>bang1mang2, help busy, 'to help'</td>
<td>10</td>
<td>shu4zou4, receive penalty, 'to have a hard time'</td>
</tr>
<tr>
<td>31</td>
<td>da2sha1, read book, 'to read; to learn or study'</td>
<td>10</td>
<td>di1tou2, lower head, 'to give up, to succumb to'</td>
</tr>
<tr>
<td>27</td>
<td>jia1ian4, see face, 'to meet'</td>
<td>9</td>
<td>gao4chuan4, report testimonial, 'to complain to the higher authorities'</td>
</tr>
<tr>
<td>23</td>
<td>shu4jian4, sleep sleep, 'to sleep'</td>
<td>8</td>
<td>zhang1zi3, open mouth, 'to talk, to ask for'</td>
</tr>
<tr>
<td>21</td>
<td>nian4sha1, read (loud) book, 'to learn, to study'</td>
<td>8</td>
<td>ting1ke4, hear class, 'to attend classes'</td>
</tr>
<tr>
<td>15</td>
<td>lai2xin4, come letter, 'to hear from'</td>
<td>8</td>
<td>feng4ming4, hold order, 'to act under orders'</td>
</tr>
<tr>
<td>13</td>
<td>xi3zao3, wash bath, 'to bathe'</td>
<td>8</td>
<td>bai4shi1, salute master, 'to acknowledge sb. as</td>
</tr>
</tbody>
</table>

To cite this output:
Table 3. Top 20 SCs in LLSCC and LCMC

Table 3 shows the top 20 SCs in the two corpora. There is a slight difference in the ordering of the top SCs in the LLSCC and the LCMC; nonetheless, the SCs overlap to a large extent. The most frequent SCs in the LLSCC are *bang1mang2* (25), *nian4shu1* (21), *jian4mian4* (18), *du2shu1* (17), *shui4jiao4* (14), *lai2xin4* (14), *hui2xin4* (10), *gao4zhu4ng2* (9), etc. The top SCs in the LCMC are *du2shu1* (14), etc. The lexical meaning of the most typical SCs seems to centre on common bodily sensations (e.g. to blush), basic everyday human activities (e.g. to sleep, to help, to meet, to shake hands, to take a shower etc.), which actually have acquired a metaphorised meaning after repeated use over time (e.g. “face red” now means “to be mad at somebody”).

The log likelihood ratio (LL = 14.49, p = 0.000) of the overall SC usage in the two major discourse modes is significant at the probability level of .001 (see Table 2). This suggests that SCs are significantly more frequent in spoken Chinese than in written Chinese. The propensity of SCs in speech is primarily realised by their repeated occurrences, because the SC types of both discourse modes are more or less the same (108 vs. 104). A further analysis of the actual SC types used in LLSCC and LCMC shows that the two datasets share the top SCs, only with slightly different ranking orders.

What is more interesting is the breakdown of the frequency distribution of SCs in the genres covered in the spoken and written corpora. Among the seven text categories within spoken Chinese, discontinuous SCs occur approximately 17 times as frequently in TV and movie scripts (646.4 occurrences per million words) as in formal debates (38.5 occurrences per million words). Interestingly enough, no single discontinuous use of SCs is found in news editorials in our written corpus. This last sub-genre in China is regarded as expressing the official viewpoints of the government or the Party. Across the 14 text categories of written Chinese which contain SC occurrences, humorous texts (865.1 occurrences per million words) have about 35 times as many SCs as academic prose (24.9 occurrences per million words).

Examining the SC occurrences in two broad discourse modes, spoken and written, and especially within the different text categories, we can discern a continuum from typical written to typical spoken genres as the frequencies of SCs increase. Both TV and movie scripts and humorous texts figure most prominently in the two genres, as one can expect, in that the two text categories are characteristic of the most dramatic, or involved, use of language. SCs seem to be a means of achieving the effect of involvement. On the other hand, academic prose and formal debates aim to convey a considerable amount of information within limited time.

4.2 Lexical and grammatical patterning of insertions

The elements between SCs and SCTs are referred to as insertions in this study. Their length varies. The corpus analysis provided the relevant calculation of the span of insertions. Intuitively it seems to be no longer than ten tokens. However, it is not the
length of the insertions but rather their lexico-grammatical properties which are of more theoretical relevance to the understanding of SC behaviours.

4.2.1 Verbal satellites of SC constructions

The first type of SCs with insertions found in our corpora are those in which the head and the tail are separated by an aspect marker. There are three subtypes of these which are given in Table 4.

<table>
<thead>
<tr>
<th>SC types (%)</th>
<th>SC tokens (%)</th>
<th>SC TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC\textsubscript{H-le} SC\textsubscript{T}</td>
<td>42 (25%)</td>
<td>74 (13%)</td>
</tr>
<tr>
<td>SC\textsubscript{H-guo} SC\textsubscript{T}</td>
<td>15 (9%)</td>
<td>22 (4%)</td>
</tr>
<tr>
<td>SC\textsubscript{H-zhe} SC\textsubscript{T}</td>
<td>12 (7%)</td>
<td>35 (6%)</td>
</tr>
<tr>
<td>Total</td>
<td>69 (42%)</td>
<td>131 (23%)</td>
</tr>
</tbody>
</table>

Table 4. Type 1a “SC\textsubscript{H} ASP SC\textsubscript{T}” constructions in LLSCC and LCMC

Of the 575 instances of SCs in the two corpora 13% (74) involved “SC\textsubscript{H-le} SC\textsubscript{T}” of which there are 42 different types. We hereby call the three-morpheme pattern “SC\textsubscript{H} ASP SC\textsubscript{T}” Type 1a, the first type of SCs with aspect marker insertion. Some examples are provided in (10).

(10) jian4-le mian4 (12), see ASP face, ‘to have met’
hong2-zhe lian3 (10), red ASP face, ‘to be feeling unhappy about sb.’

Among the most common SC\textsubscript{H} ASP SC\textsubscript{T} constructions, some (e.g. tie3-le xin1, iron ASP heart, ‘to be determined’, and shui4-guo jiao4, sleep ASP sleep, ‘to have slept’) are highly idiomatic in that they are more readily accessible to native speakers than they are without an aspect marker. SC constructions with an aspect part can be longer than three morphemes. These SCs with expandable insertions together with type 1a SCs are labelled “Type 1b” in the discussion below. E.g.

(12) 突然 来了封信 叫我…(from LLSCC)

Tu1ran2 lai2-le feng1xin4 jiao4 wo3 …
Suddenly come ASP CL letter make me...
Quite out of blue, a letter came, and made me...

<table>
<thead>
<tr>
<th>SC types (%)</th>
<th>SC tokens (%)</th>
<th>SC TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC\textsubscript{H (?)}ASP (? SC\textsubscript{T}</td>
<td>91 (55%)</td>
<td>244 (43%)</td>
</tr>
</tbody>
</table>

Table 5. Type 1b “SC\textsubscript{H (?)} ASP (? SC\textsubscript{T}” constructions in LLSCC and LCMC

In terms of types, namely distinct word forms, of SCs, over half of them (55%) contain an aspect marker, either -le (perfective or more accurately “actual” aspect, cf. Xiao and McEnery, 2004), -guo (experiential aspect) or -zhe (durative aspect). Moreover, SCs with an interposed actual aspect marker -le constitute one quarter (25%) of Type 1
SCs. Therefore, it is safe to say that a typical SC bears an aspect marker, and the prototypical grammatical pattern of SC is $SC_{T} \rightarrow SC_{T}$, a three-morpheme verbal construction with an actual aspect marker in the middle. The length of Chinese words is in reverse proportion to their relative token frequency (see Xiao, Rayson and McEnery, 2009, pp. 13-14). Back to our case of SC, although there is a good deal of variation of insertion types and combinations other than the three-morpheme SCs, they only account for a fractional part of all SCs.

In addition to aspect markers, quite a few other elements can be attached to the SCHs. Some typical instances of these SCs include resultative verb complements (RVCs), e.g. $nian4 \ wan2 \ shu1$, ‘to finish school’ (4), $jing4 \ xia4 \ xin1$, ‘to calm one’s mind’ (3), and $xi3 \ wan2 \ zao3$, ‘to finish one’s shower’ (2). The single RVC in the middle includes words like $wan2$, ‘over, finished’ (10/26), $xia4$, ‘down, downward’ (3/26), $shang4$, ‘up, upward’ (3/26), $ban3$, ‘well, done’ (2/26), etc. The majority of the 20 out of the 26 uses of RVCs expresses a more or less perfective sense, which is hardly surprising, given that RVCs can be analysed as markers of the “completive aspect” in Chinese (Xiao and McEnery, 2004).

<table>
<thead>
<tr>
<th>SC type (%)</th>
<th>SC token (%)</th>
<th>SC TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SC_{T} \rightarrow RVC \ SC_{T}$</td>
<td>20 (12%)</td>
<td>26 (5%)</td>
</tr>
</tbody>
</table>

Table 6. Type 2a “$SC_{T} \rightarrow RVC \ SC_{T}$” insertions

The 0.77 type/token ratio in Table 6 suggests that the $SC_{T} \rightarrow RVC \ SC_{T}$ type is quite productive, much more flexible than the $SC_{T} \rightarrow ASP \ SC_{T}$ type (TTR, 0.53) or the $SC_{T} \rightarrow ASP \ SC_{T}$ type (TTR, 0.37). However, the semantic analysis of RVCs reveals that the diversity of RVCs plays a similar role in making aspectual modification; most often it adds perfective meaning to the verbal construction.

<table>
<thead>
<tr>
<th>SC types (%)</th>
<th>SC tokens (%)</th>
<th>SC TTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SC_{T} \rightarrow RVC \ SC_{T}$</td>
<td>20 (12%)</td>
<td>66 (12%)</td>
</tr>
</tbody>
</table>

Table 7. Type 2b “$SC_{T} \rightarrow RVC \ SC_{T}$” insertions

To sum up, the major types of elements attracted by the verbal heads of SCs are aspect markers and resultative verb complements, which are sometimes regarded as special types of aspect markers (cf. Xiao and McEnery, 2004). These elements float around verbal heads modifying their telicity, progress, etc. The SC instances are characteristic of another big cluster of insertion elements which are more often found before SCs, namely the nominal/complement components. Most typical items in this category are quantificational expressions, classifiers, and various forms of pre-modifiers.

4.2.2 Nominal/Complement satellites of SC constructions

**Quantifier**

There are 108 instances of quantificational expressions in the insertion structures (19% of all SCs). Among them, 67 (62% of quantifiers) are $yi1$, ‘a, one’. The quantifiers
can be grouped as either approximate quantificational expressions (e.g. ‘several’; ‘a couple of’) or exact numbers (e.g. ‘a’, ‘one’; ‘three’). In Chinese, especially in everyday language, ‘two’, can mean two to three, or “a couple of”, when it is unstressed.

Classifier
Nominals in Chinese are typically preceded by a classifier. In our data, 116 SCs (21% of all SCs) contain a classifier. The most common classifier is which occurs 40 times and accounts for 34% of all the quantifiers among the split SCs.

The combination of quantifier and classifier, and occasionally demonstratives in the place of a quantifier or preceding a quantifier, result in a structure similar to the determiner system in English. Actually, quantificational expressions and classifiers either specify number, frequency, etc. or vary categorically the statuses or attributes of the ensuing nominal elements.

Modifier
The modifier category here means the pre-modifying element(s) of the nominal component of an SC, which typically includes adjectival modifiers (63 times, 11% of all SCs), nominal items (59 times, 10% of all SCs), possessive personal pronouns (64 times, 11% of all SCs), question words (i.e. ‘what’, etc. 26 times, 5% of all SCs), and also combinations of these elements.

There can be in an SC structure more than one adjective, nominal item or personal pronoun either repeated or in juxtaposition. Very often (10 times in our data) a possessive marker is used between the modifier and SC. In authentic data, it is not surprising to see repetition, false starts or swear words in insertions. These residual categories were not discussed.

4.2.3 Summary of insertions in SC constructions
On the basis of the above comprehensive categorisation of the insertions within discontinuous uses of SCs, we can now summarise their structural composition. From a syntagmatic perspective, a typical SC in use follows the pattern, or the colligation (Firth, 1957/1968, p. 181) of Chinese SCs:

\[
\text{SC}_H + \text{NEG} + \text{ASP/RVC} + \text{MC} + \text{CL} + \text{MOD} + \text{SC}_T
\]

Fig. 1. A syntagmatic pattern of typical SCs

Paradigmatically, the 166 instances of the “SC$_H$...SC$_T$” type in our data are potential candidates for an SC template. Each grammatical slot can be filled by a limited set of morphemes, as detailed in previous sections. Figure 4 depicts fully developed discontinuous usage, but obviously in reality SCs are realised in different forms of different length and different combinations. In the figure, SC$_H$ is the verbal head and SC$_T$ the nominal/complement element of an SC. The “SC$_H$...SC$_T$” can be
interrupted by one single element of the 4-5 categories between the two parts of the discontinuous construction. For instance, there can be a negator (NEG), an aspect marker (ASP), a quantificational expression (MC), a classifier (CL), different forms of pre-modifiers. Alternatively, the insertion of a discontinuous SC can be a combination of two or more elements listed in Figure 4. Figure 5, however, can be seen as a better illustration of the typical lexical and grammatical patterning of discontinuous SCs, allowing for all types of selection and combination. Indeed, most discontinuous SCs tend to be short and gravitate to the verbal heads. In our data, 275 out of 566 (49%) of the discontinuous SCs are three-morpheme constructions. Moreover, 42% of the three-morpheme discontinuous SCs are with just one aspect marker, with 25% containing -le as the single insertion.

Fig. 2. Typical lexical and grammatical patterning of discontinuous SCs

The diagram gives a skeletal but explanatory description of the lexical and grammatical patterning of SCs based on the two million words of spoken and written data. The diagram is read from left to right, starting from the verbal element SC_H and heading for the other component of the SC, namely the SC_T.

The nodes given in the balls (except those for SC_H and SC_T) stand for the most frequent lexical item of the grammatical categories aspect marker, resultative complement, quantifier, classifier, and modifier. The ± symbol alongside the “aspect marker” and the “resultative verb complement” shows that they can receive a preposed negator, most likely bu4, ‘not’. The pronunciation of -le, however, changes to liao3, when bu4 precedes it. Actually, liao3 is more of a verb complement instead of an aspect marker, which in our discussion the two categories serve perfective aspect marking purposes.

4.3 Contextual clues of SC discontinuity
If the lexical and grammatical investigation of SC insertions reveals the internal configuration of SCs, the lexical semantics of $SC_{1}$'s and the $SC_{2}$'s and their co-text also deserve special attention. This section explores the interaction of the contextual clues with the split use of SCs. In our manual annotation, three values, neutral, negative and positive, were marked up for the “wordsemantics” of the templatic units comprised of $SC_{1}$'s and $SC_{2}$'s. The results show that 451 SC templates (80% of all SCs) are neutral, 35 (6% of all SCs) positive, 72 (13% of all SCs) negative, and there are some residual cases which are hard to determine. Therefore, most SC templates are neutral in their connotation.

In addition, we annotated sentence types as declarative, negative, exclamatory, interrogative, rhetorical or imperative. The sentence type statistics indicate that there is no significant correlation between sentence types and the discontinuous use of SCs.

In summary, both the internal and external structure analysis of discontinuous SCs suggests that the split SC usage is not sensitive to its sentential context in general. Rather, the structures and functions of the insertions within SCs have more weight in keeping $SC_{1}$'s and $SC_{2}$'s apart.

4.4 Discourse pragmatic properties of SCs and their insertions

The propensity of SCs in speech has already shed some light on the discourse preference of SCs in speech. To make sure about the correlation of split SCs to spoken Chinese, the combined uses of the 166 SCs were searched once again in LLSCC and LCMC. 1494 and 1577 occurrences of the 166 SCs in the two million word texts do not show statistical difference. That is to say, the combined uses of SCs are evenly distributed in both spoken and written texts. However, in the written data, news review (22 times, 408 times per million words, 1.6% of the written corpus) and academic prose (87 times, 542 times per million words, 2.1% of the written corpus), two most formal sub-genres are the two text types which contain least combined uses of SCs. At the other end of written genre cline, mystery and detective stories (135 times, 2,788 times per million words, 11% of the written corpus), and humorous texts (61 times, 3,298 times per million words, 13% of the written corpus) are two text types which contain most combined uses. Therefore, both split use and non-split use of SCs point to the spoken genre preference of SC usage, and the split usage of SCs is especially so.

Statistics indicate that the highly grammaticalised $yi$ and $ge$, which either appear on their own or in the combination of the two, are two prominent items that go with SC's. According to Biq (2004, p. 1663), $V_{yi}ge_{N}$ and its variant $V_{ge}N$ encode a sense of trivialness, casualness or unremarkableness. Biq (2004, p. 1660) even argues that the “$V_{1}(yi)ge_{V_{2}}$ construction provides a ‘buffer zone’ in the on-line production situation”. The classifier $(yi)ge$, together with other adjectival and nominal modifiers mitigate, soften, or qualify the properties of the nominal elements.

4.5. The defining principles of SCs

Based on a corpus analysis, we have outlined the prototypical lexical, grammatical patterning and discourse interpretation of SCs. However, we have yet to
address the issue which has been at the heart of the controversy surrounding SCs, namely the morpho-syntactic status of SCs.

Unlike the scripts of other alphabetic languages, Chinese characters, the basic orthographic units, are “bearers of basic semantic and grammatical content” (Wang, 1973, p. 59). We would like to suggest some key canonical properties of SCs in relation to their grammatical and phonological characteristics.

In the previous literature, the dispute surrounding SCs has centred on the morpho-syntactic status of the split use of SCs. Those who accept the gestalt status of SCs in their lexical semantics argue that split SCs are allomorphs of their isomorphic combined uses. Structural account of SCs, however, maintains that to be a word lexical integrity has to be observed. So this latter line of argument favours the phrasal status of split SCs and lexical status of combined SCs. But what this latter treatment fails to explain is the high boundedness or great idiomatisation of the two SC constituents, i.e. many SC_{i}s are always to be found after certain SC_{ii} verbal heads.

We would here follow McCarthy and Prince (1993) and Feng (2001, 2002) in taking a prosodic basis for the compound vs. phrase distinction in Chinese, drawing on our corpus analysis. We share Feng’s claim (2002, p. 134) that the basic compound in Chinese has the prosodic structure of a Prosodic Word (PrWd), and that basic compounds are two or three syllables long due to the Foot Formation Rule (FFR), with disyllabicity as its predominant form. Thus the combined uses of SCs are compounds. To complement Feng’s account of discontinuous SCs, we would take his FFR as our starting point, and propose, on the basis of our analysis of two million words of spoken and written data, the following typology of the manifestations of SCs:

a) Combined uses of SC_{i}s and SC_{j}s were argued to be compounds due to the FFR;

b) SC_{i}s and SC_{j}s separated by one single aspect marker were considered as compounds according to the so-called Mending Device of Trisyllabic Foot (ibid., p. 112), given that an aspect marker may be viewed as the near-inflectional suffix to the preceding verb; while

c) SC_{i}s and SC_{j}s separated by other grammatical categories like numerals, quantifiers, adjectival modifiers, etc. were seen as phrases.

Our quantitative data explicate that over half or more (if RVCs are seen as quasi-aspect markers) of split uses of SCs, together with their continuous cognates, can be analysed as legitimate compounds. The rest are phrases. To make our proposal more accessible to similar discussions on SCs, we would advance two sets of continuum type of criteria for identifying SCs in Chinese.

A) The structural criteria: Host dependency: head dependence > tail dependence

The host dependency criterion (a > b > c) of the canonical approach perceives:

a) SCs with a clitic-like aspect marker (e.g. the perfective marker -le) as compounds instead of phrases;

b) SCs with resultative verb complements attached to the main verb as quasi-compounds; and
c) other modifiers (classifiers, modifiers, etc.) attached to SCCs, represented typically by a noun or complement, as least possibly compounds.

B) The phonological criteria: PrWd restriction
We propose that the various manifestations of SCs define a continuum of phonological conditions as a complement to the grammatical criteria (a > b > c):

a) The combined uses of SCCs and SCCs are disyllabic compounds;
b) SCs in which the SCCs and SCCs are separated by one single morpheme under the Trisyllabic Foot Rule are possible compounds; while
c) SCC and SCC separated by multi-syllable units in the form or combination of quantifiers, adjectival modifiers, etc. are phrases.

Both the structural and phonological criteria need to be considered in determining the wordhood of a candidate SC. When there is a mismatch between a grammatical word and a phonological word, we would give priority to the host dependency criteria.

5) Activities
We have shared our research findings by giving presentations at national and international conferences.


6) Outputs
6.1 Publications
Our corpus-based empirical work has led to the following list of publications.

(Forthcoming) Bang-le yi ge da mang (offered a big helping hand): A corpus study of the splittable compounds in spoken and written Chinese. Language Sciences.


6.2 Software tools
Xu, Jiajin & Yunlong Jia (2009). Collocator 1.0: A collocation extraction tool.

7) Impacts
In addition to the above, we have shared our research findings with many Chinese linguists and conducted discussions on the online corpus linguistics forum (http://www.corpus4u.org) run by two of the researchers of the project.

8) Future research priorities
The project has lead to the PI extending her research to other aspects of Chinese morpho-syntax, especially passive constructions in different varieties of Chinese, as reflected in her recent key note address at the 17th International meeting of the Association of Chinese Linguistics in Paris.

(Word count: 4,998 words)