Available online at www.sciencedirect.com
ScienceDirect
Language Sciences 32 (2010) 464–487
Language Sciences
www.elsevier.com/locate/langsci

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

Anna Siewierska [a], Jiajin Xu [b,*], Richard Xiao [c]

[a] *Department of Linguistics and English Language, Lancaster University, LA1 4YT, United Kingdom*
[b] *National Research Centre for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089, China*
[c] *Department of English and History, Edge Hill University, Ormskirk, Lancashire, L39 4QP, United Kingdom*

Received 20 January 2009; received in revised form 4 July 2009; accepted 12 August 2009

Abstract

Splittable compounds (SCs) are verbal constructions in Chinese that consist of two parts which are separable by some interposing elements, though they behave like and are usually considered as single words when they are not separated. For many years, SCs in Chinese have presented a challenge to existing morphological and syntactic theories for their morpho-syntactic status. The present study takes a corpus-based approach to the SCs in their interaction with morphology, syntax, and pragmatics, aiming at producing a systematic and realistic account of SCs as attested in 2 million words of authentic spoken and written Chinese data. The results show that the typical grammatical pattern of SCs is constitutive of an aspect marker (*-le*, *-zhe*, *-guo*) or resultative verb complements as post-verbal adjacent elements (54% of all SCs), and a quantifier, a classifier, a modifier or a combination of two or more of them which precede the nominal components of SCs. Drawing on morpho-syntactic and phonological criteria, the split uses, together with their combined uses, of SCs with one inserted aspectual morpheme are viewed as words, while the others are regarded as phrases. From a discourse-pragmatic perspective, the split use of SCs is more often found in the spoken genres of Chinese. Insertions of SCs tend to function as mitigation or modification to the verbal heads or final nominal/complement elements.

Keywords: Chinese morphology; Corpus methodology; Grammatical word; Phonological word; Splittable compounds (SCs)

Abbreviations: ADV, adverb; ASP, aspect marker; AUX, auxiliary; CL, classifier; INT, interjection; LCMC, Lancaster Corpus of Mandarin Chinese; LFG, Lexical Functional Grammar; LIH, Lexical Integrity Hypothesis; LL, log likelihood ratio; LLSCC, Lancaster Los Angeles Spoken Chinese Corpus; MC, quantifier; MOD, modifier, modification; NEG, negator; NLP, natural language processing; NP, nominal phrase; PART, particle; PrWd, prosodic word; PSC, Phrase Structure Condition; RVC, resultative verb complement; SC, splittable compound; SC_H, the verbal head of an SC; SC_T, the nominal/complement (tail) component of an SC; TTR, type token ratio.

* Corresponding author. Tel.: +86 10 8881 6828; fax: +86 10 8881 0376.
E-mail address: xujiajin@bfsu.edu.cn (J. Xu).

0388-0001/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.langsci.2009.08.002

1. Introduction

Splittable compounds (SCs hereafter), which are better known among Chinese linguists as *Lihe Ci* ('Separate–combined Words') in Chinese morphology, are hardly ever an undisputed issue in view of their controversial morpho-syntactic status. The controversy over SCs stems, on the one hand, from discrepancies in regard to what constitutes a morphological word in Chinese, and on the other, from the gestalt status of lexical semantics with flexible morpho-syntactic variation.

SCs are very frequent in modern Chinese. According to Shi (1999, p. 123), there are 2960 SC entries in the Modern Chinese Dictionary (1996 edition),[1] which accounts for 51.63% of all verb object words in the dictionary. The 2002 edition of the dictionary updated the SC family to 3236. Among the four subtypes of SCs, verb object compounds constitute 97% of the category, while the other three categories combined (i.e. verb complement compounds, subject predicate compounds, and coordinative compounds) make up the remaining 3% (Zhu, 2006, p. 29). Quite clearly thus, the verb object type is the predominant form of SCs in Chinese. It needs to be mentioned though that the categorisation and marking up of SCs in the dictionary by its compilers, who are leading grammarians, are by no means uncontroversial and have in fact been the object of much criticism.

In order to facilitate discussion, we will first present a characterisation of SCs which is purely descriptive. Like standard verb object compounds in English (e.g. *pickpocket*), German (e.g. *Schreibtisch*, 'writing-desk') or Italian (e.g. *lavapiatti*, 'washing machine'), Chinese SCs consist, in the main, of two parts, the head (SC_H) and the tail (SC_T), which can be used as separate lexemes but can 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]
(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*, as in (3).

(3) ***xin1*** *yi4zhi2* ***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. Moreover, from a syntactic point of view, constructions like *dan1xin1* can take objects, i.e. they may function as a transitive verb. We see in (4) that *dan1xin1* can take a direct object, and in (5) that it can even take a subordinate clause.

(4)	*Wo3men*	***dan1xin1***	*ta1*	*de*	*jian4kang1*.
	we	worry	s/he	AUX	health
	'We worry about his/her health.'				

[1] Modern Chinese Dictionary (*Xiandai Hanyu Cidian*) is the most authoritative Chinese dictionary in China, whose status is similar to Oxford English Dictionary in the English world. The dictionary uses double slash "//", as in *xi3//zao3* and *dan1//xin1*, to represent the splittability of such units, meaning that they can be interposed by other elements. Namely, the instances of "//"s are equal to the SCs accepted by the compilers of the dictionary.

[2] In this paper, when an SC example is provided, the Romanised pinyin gloss in italics, with the tone indicated by an Arabic numeral (and the absence of an Arabic numeral either suggests a neutral tone or a generic use of the morpheme), is used to represent Chinese SCs in characters. Chinese characters are also given where necessary. A Chinese example of SC is normally followed by a literal gloss and then the corresponding English translation in single quotes. In this paper, examples 1–5 and 8 are based on our own introspection to show what an SC typically is and what is not; all the other examples are cited from our corpus data.

(5) *Da4jia1* ***dan1xin1*** *ta1* *neng2* *bu4neng2* *na2dao4* *qian1zheng4*.
everyone worry s/he can or not get visa
'Everyone worries whether s/he will be able to get the visa.'

dan1xin1 is not exceptional in its ability to take a direct object. There are quite a few SCs which do so, though SCs used as intransitive verbs are more common (Li and Thompson, 1981, p. 76).

Confronted with special types of morpho-syntactic items such as *dan1xin1*, 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; Packard, 2003; Zhou, 2006, to name but a few). In this paper we seek to resolve some of the controversies surrounding SCs in Chinese by providing a detailed corpus-based investigation of their distribution, structure and functions.

The paper is organised as follows: In Section 2, we will take a closer look at the "wordhood" as applied to Chinese and previous discussions of SCs that the different views of what constitutes a word in Chinese have engendered. In Section 3, we will discuss, in some detail, our corpus-based research design of the lexico-grammatical behaviours of SCs. The corpus data as well as annotation and computational and qualitative analytical procedures will also be presented and explained. Then in Section 4, the prototypical internal structure, sentential and discoursal contexts and properties of SCs will be explored. Finally, in Section 5, a structural and phonological delimitation of the morpho-syntactic status of SCs is proposed on the basis of the previous analyses.

2. Words, compounds and SCs in Chinese

2.1. Chinese words: a historical account

Classical Chinese, especially pre-Qin (earlier than 221 B.C.) Chinese, is well documented for its monosyllabicity (Branner, 2003, p. 49; Lü, 1961). Etymologically, most Chinese morphemes are monosyllabic. There are only very few disyllabic morphemes, traditionally called *Lianmian Ci* (i.e. disyllabic alliterations, like *fang3fu2*, 'seemingly'; and disyllabic rhymes, like *hun2dun4*, 'chaos'), as well as a number of early loan words (e.g. *bo1li*, 'glass', *pu2tao*, 'grape', etc.). The monosyllabicity is mapped onto the Chinese writing system by single syllable words, each having their own meaning. This etymological heritage has been a most plausible justification and incentive for a new "character-centred approach" (see also discussion on "sinogram" in Section 2.3.2) to Chinese morpho-syntax (Xu, 1997; Pan, 2006; Zhou, 2006). This approach assumes that historically Chinese morphology used to involve nothing more than individual distinct morphemes (viz. characters in writing), and syntax was the juxtaposition of individual characters.

2.2. Compounds and compounding of Chinese

There has been a gradual increase in syllabicity and polysyllabicity in modern Chinese, which is most likely to be motivated by the growing need to express complex ideas, so that the overwhelming majority of words in modern Chinese are disyllabic in terms of word types (cf. Xiao et al., 2009, pp. 12–13). This change, however, cannot be recognised from the orthography, given that the Chinese writing system does not use white spaces to separate words in running texts. Even the use of punctuation in writing is a quite recent practice initiated to facilitate reading. In speech, native speakers are capable of telling apart one word from another with the aid of cues such as pauses and rhythmic patterns. Moreover, the psychological reality of the conceptual construct of Chinese words in native speakers' minds is assumed to help to retrieve meaningful units from the stream of sounds, or from the string of characters.

The historical development of Chinese words from monosyllabicity to disyllabicity (and less often to polysyllabicity) sees the breakup of the one-to-one correspondence of syllable, morpheme and word(hood). However, the transformative side is the flexible length as well as productivity of some more grammaticalised morphemes. Some words become bound roots, and others function words.

The general lengthening of lexical units is likely to be the result of the "habitual" (in Firth's (1951/1957, p. 179) sense of the term) combination of two or more morphemes. Compounding has been the most "productive" and "widespread" means of word formation in modern Chinese (Ceccagno and Basciano, 2007, p. 208; Yang, 2003, p. 134), and it continues to be an important way of creating new words in Chinese. In fact, Haspelmath (1992, p. 71, cited in Packard, 2003, p. 262) considers "compounding in Chinese as analogous to grammaticisation" as the second member attached to the head involves "semantic generalisation and phonological erosion . . . [which] is clearly a completely analogous diachronic process". Moreover, a generally acknowledged explanation (as argued by Wang, 1989, p. 2) for the underlying motivation of the proliferation of compounding in the history of Chinese might be the need of disambiguation of the great number of homophones in Chinese. Compounding therefore becomes a handy means of making words more distinctive and specific.

Over the centuries, compounding has too brought significant changes into Chinese morpho-syntax. According to Zhang (2007, p. 4), over two-thirds of the Chinese lexicon are disyllabic though, in term of word tokens, monosyllabic words account for more than half of running texts (Xiao et al., 2009, p. 13). This suggests that about half of the modern Chinese lexicon are disyllabic. As stated at the beginning of the paper, about 3000 of such compounds are SCs. The fact that so many verbal compounds are splittable forces linguists to redefine what a word or compound is in Chinese, and, above all "forces us to establish criteria for distinguishing verb–object compounds from verb–object phrases" (Li and Thompson, 1981, p. 73).

In the light of the above, no grammar (or account of morpho-syntax *per se*) of Mandarin Chinese can turn a blind eye to the "verb–object paradox"[3] (Packard, 2003, p. 108), because SCs are by no means a marginal morphological phenomenon. In the literature we have noted discussions on SCs in: (1) the structuralist studies informed by Western grammars; (2) the sinogram-based theory developed by Chinese linguists; (3) the formal principles (in the generative tradition and Lxical Functional Grammar); (4) the phonological interaction with morpho-syntax; and (5) methodologically, the computer-assisted quantitative approach to SCs. These will be briefly reviewed respectively in the next section.

2.3. Previous analyses of Chinese SCs

2.3.1. Structuralist analyses of SCs in mainland China

Although SCs are called *Lihe Ci*, 'Separate–combined Words', in Chinese, very few linguists in China would assume, without hesitation, that SCs are simply words which may be split. The predominant view is that SCs are words when the verbal and nominal/complement morphemes appear together, and are phrases when used separately (Lu, 1957; Zhu, 1982, among others). Lü (1979), by contrast, advances a more prudent view, namely that we should recognise that there are "middle-state" or transitional categories in morpho-syntax. The contentious status of SCs lies in the conflicting criteria—the gestalt lexical semantics criterion for lexical status and the lexical integrity of different parts of the structure. Similar views were expounded by Chao (1968) with his oft-cited five conditions and the "ionisation" view of verbal compounds in Chinese, spoken Chinese in particular.

Chao (1968, pp. 415–480) suggests the following defining criteria (Chao, 1968, p. 415) for "infixable" (Chao, 1968, p. 437) and "expandable" (Chao, 1968, p. 438) compounds (both verb–object (V–O) and verb–complement (V–R) types): (1) one or both of the constituents are bound; (2) the object has a neutral tone; (3) the construction as a whole is exocentric; (4) the meaning is (specialised) lexical; and (5) the constituents are inseparable. Chao argues that a V–O construction has to satisfy one or more of the above conditions to be classified as a compound. The five conditions are, however, not unproblematic. Li and Thompson (1981, pp. 73–81), for example, reject the neutral tone condition. Other scholars such as Huang (1984) and Yu (2003) argue against each of the five conditions. Their general objections are that the conditions are too vague to be operationalised in identifying a compound, and that the conditions may conflict with each other and that the incorporation of infixable and expandable compounds violates the Lexical Integrity Hypothesis (see Section 2.3.3).

[3] Since verb object compounds constitute 97% of SCs (Zhu, 2006, p. 29), in the literature very often scholars are more interested in V–O compounds alone.

Considerably more interesting and tenable than Chao's five criteria for what constitutes a compound in Chinese is his notion of "ionisation" (1968, p. 159), referring to the discontinuous behaviour of many V–O and V–C compounds. This idea, initially advanced during the 1930s, is similar to Firth's (1951/1957) notion of collocation. Ionisation likens the "floating around" of morphemes of an expandable V–O or V–C compounds within the same sentence or close context to certain ions as "a chemical compound floating around in the same solution with its partners". The separated morphemes in the compounds are "ionised" or "ionisable" (Chao, 1968, p. 159). The stable separate state or the discontinuity is sustained by a certain inherent collocability, similar to the physical–chemical properties that keep the ions of a solution in a stable separate state. This account allows for the separability of the compounds and at the same time identifies the "similarity" or unnamed force that keeps the morphemes not far from each other.

2.3.2. Sinogram-based approach to SCs

The dissatisfaction of Chinese linguists with the Western concept of a morphological word leads to the introduction of the "sinogram" as the minimal grammatical unit for Chinese (Xu, 1997). This sinogram-based theory views Chinese characters as the psychologically real minimal unit of Chinese. A sinogram, a composite constituent of sound, form and meaning, is formally equal to a character in written Chinese; the term "character", however, is not used because it is considered to be orthographical and not a good grammatical term.

The sinogram-based approach to SCs has been recently adopted by Zhou (2006) in an attempt to explain the nature of SCs. Zhou holds that sinograms are the sole legitimate building blocks of Chinese, and all larger units in Chinese are sinogram clusters of some sort. This viewpoint disregards the boundedness of some sinograms which are only found in combination with certain other sinograms, but never used alone. Nevertheless, the sinogram-based approach to compounds suffers from an over-reductionism to morpheme-like units which does not allow for a middle-ground between morpheme and syntax.

2.3.3. Formal representations of SCs

Working within a Chomskyan framework, Huang (1984) seeks to delimit the nature of Chinese compounds, verb object compounds and resultative compounds (viz. verb complement compounds) with reference to two formal principles, the Phrase Structure Condition (PSC) and the Lexical Integrity Hypothesis (LIH). The PSC, in simplistic terms, allows one and only one constituent following the verbal head in a given Chinese sentence, though a verb can be preceded by an indefinite number of constituents (including subject and adverbial modifiers) (Huang, 1984, p. 54). The LIH dismisses the possibility of any phrase-level rule that may affect the subpart of a word (Huang, 1984, p. 60). Bearing in mind the generative and generative semantics conditions, Huang acknowledges the dual status of SCs (Huang, 1984, pp. 68–70), and suggests three possible ways of pigeonholing SCs: (a) to list all SCs in the lexicon as both words and phrases; (b) to list them as words in the first place; or (c) to list all SCs as phrases. Essentially, the X-bar theory of phrase structure will invoke reanalysis rules to identify the lexical and phrasal status of an SC.

Another formal treatment of SCs is provided by Yu (2003) in his analysis of Mandarin and Cantonese discontinuous V–O compounds. This lengthy in-depth analysis is carried out in the context of Lexical Functional Grammar (LFG). Yu considers the lexical, syntactic and semantic properties (or "specifications" in Yu's terms) of V–O compounds and argues that the correspondences between a-structure (argument structure) and f-structure (functional structure), and between f-structure and c-structure (constituent structure) in LFG are of help in identifying V–O compounds. However, the mapping of properties onto the three planes allows for a careful underpinning of SCs but fails to take account of some of their formal and semantic aspects.

2.3.4. Phonological considerations of SCs

As pointed out by Mathews (1991, p. 209), "the word tends to be a unit of phonology as well as grammar". It is, therefore, to be expected that in addition to the morpho-syntactic accounts of SCs, there are also phonologically based ones. In fact, phonological considerations have figured prominently in recent efforts to account for the paradoxical nature of words and phrases in Chinese. Given the problems inherent in defining the wordhood of SCs in Chinese, Feng (2001) turns to the framework of prosodic morphology (McCarthy and Prince, 1993, 1995), arguing that the notion of a prosodic word may be more suitable and practical for Chinese

than the traditional "word". In prosodic morphology, a prosodic word is realised by a prosodic foot, which is in turn a combination of syllables which are composed of mora. Chinese, a syllable-timed language, is characterised by the presence of metrical patterns of disyllabic or trisyllabic feet. The prosodic-morphological approach to Chinese words avoids the problem of the double classification of SCs as words on the one hand, and phrases on the other. We will return to this approach later in our discussions based on the data retrieved from spoken and written corpora.

2.3.5. The computational approach to SCs

There is a large body of literature in computational linguistics which focuses on the word segmentation of Chinese running texts, of which the identification of discontinuous constituents is a small but important aspect. In this section, however, we are concerned only with the linguistically-focused corpus analyses of SCs.

A well-designed corpus-based investigation of SCs is presented in Wang (2001). Although the goal of the research is the auto-identification and mark-up of any input texts either for NLP and language study or pedagogic purposes, the study also looks into some local contextual properties of SCs and the variation in the behaviour of SC across different genres (e.g. news, fiction and play). Wang starts with a base list of 3877 SCs (*as per* Yang (1995) with supplements from other references) to search the 5,150,880 characters (approximately 3 million words if tokenised) for candidate SCs. The patterns of erroneous hits of SCs are summarised, and the edited clean sentences with SCs are analysed for insertion patterns. After closed and open tests of automatic identification and mark-up drawing on the results from the previous analyses, Wang achieves a precision rate of 81.74% with a recall of 98.27%. Her research is noteworthy in that it identifies the problematic cases and general patterning of SCs found in authentic texts, which is most useful for future computational analysis of SCs. The major problem with the research lies in the corpus used, which consists largely of fiction, news reportage but does not include any spoken data, though SCs are much more common in speech (see also Section 4.1).

Ren and Wang (2005) investigate 423 SCs in a corpus of fiction consisting of 13.7 million characters. Both continuous and discontinuous uses of the 423 SCs are extracted. But unlike Wang (2001), this study does not seem to involve human filtering of noise hits which could be a defect in the procedure.

2.3.6. An interim summary

We have seen so far that there is a plethora of views on how to deal with Chinese SCs, the problematic nature of which springs from the fact that in the Western linguistic tradition, words cannot be discontinuous. The analyses that have been proposed may be summarised in simple componential terms, as diagrammed in Fig. 1, as involving blending (or compounding) of two component morphemes, or separation of composite compounds, or insertion (or infixation) between the initial and final elements. These two key elements can be joined by their inherent or idiomatic core meaning, provided that an appeal to semantics is allowed in determining a morpho-syntactic structure.

If overall semantic content is accepted as a criterion, then the resilient shape of SCs can be well accommodated as a "grammatical construction" in the functional linguistic tradition, so long as some internal and external features of the construction are identified, which calls for empirical evidence to lead to more plausible generalisations about SCs. Before presenting our corpus-based research design, we will explain why a corpus approach is adopted to deal with the SC issue. First of all, we opted for a corpus-based approach to SCs because corpus data have been increasingly recognised as an important resource in linguistic research and are nowadays used in nearly all branches of linguistics (see McEnery et al., 2006). Nonetheless in order to identify the SCs in our corpus and especially categorise them in linguistic terms, we need to apply qualitative

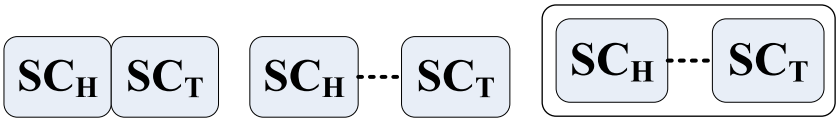

Fig. 1. Divergent views of the structure of SCs

analysis. Our research design combines both quantitative and qualitative analyses to address the following research questions regarding the linguistic behaviours of SCs:

(1) What are the common types of insertions and what are their syntactic functions in the discontinuous use of SCs?
(2) In what contexts are SCs typically used in discontinuous form?
(3) Is there a correlation between the separate use of SCs and clause or sentence type?
(4) What are the discourse functions and pragmatic meanings of SCs?

Most of the answers to the four questions will be investigated with a combined quantitative–qualitative method, such as questions 1, 2, and 3. Question 4 concerns the general tendency of SCs or variation across different genres or sub-genres, which can be accounted for largely by statistical measures. The corpus resources and analytical procedures will be presented in the following section.

3. Data and method

3.1. The corpora

Two corpora will be used in this study: the *Lancaster Los Angeles Corpus of Spoken Chinese* (LLSCC) for spoken Chinese and the *Lancaster Corpus of Mandarin Chinese* (LCMC) for written Chinese. The LLSCC comprises 1 million words of dialogues (55%) and monologues (45%) in Chinese. These represent both spontaneous (57%) and scripted (43%) speech. The LCMC is a balanced corpus of written Chinese composed of 1 million words proportionally sampled from fifteen genres ranging from news, fiction to academic prose published in mainland China within two years around 1991 (see McEnery et al., 2003). Both the LLSCC and the LCMC are marked up in XML, indicating genre information in the corpus header, as well as boundaries for paragraphs, sentences and word tokens in the body of a corpus file. The two corpora are also tokenised and annotated with part-of-speech information. They form the empirical basis for our quantitative and qualitative analysis of SCs.

3.2. The procedures

We seek to achieve our research objectives by undertaking the following procedures. First of all, we will use the 1738 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. As reviewed earlier, Wang (2001) starts off with 3877 SCs listed in the appendix of Yang's (1995, pp. 949–1018) dictionary, with a few additions from other sources. Her corpus search revealed that within her 5 million character corpus, only 1124 SC types were found, and a large number of these occurred only once. Of the 1124 SCs only 650 were found more than four times (Wang, 2001, p. 7). In view of these findings, we chose for our investigation the 1738 core SCs included in the main body of Yang's dictionary. This 1738 item list covers the most frequently discussed SCs in the literature and we are confident that these SCs will suffice for the purpose of general profiling of the SC usage.[4] The results of our corpus search confirm that only 166 SC types out of the 1738 SCs were found in the 2 million words (see Appendix II for the list of the 166 SC types). This result shows that only about one tenth of the 1738 SCs listed in Yang's dictionary are frequently used in our corpora. In retrospect, this suggests that it was wise to start with a more comprehensive list, instead of a handful of brainstormed words, in order to get a fuller picture of SC usage. On the other hand, the 1738 entries in the dictionary have to be reconsidered against empirical evidence from balanced corpus data as good instantiations of SC in Chinese. Potential SCs collected from the literature and the compilers' introspective knowledge are helpful in general discussions; yet, SC items from

[4] There are purely corpus-driven methods (e.g. concgramming, cf. Cheng et al., 2006) through which long distance dependency patterns can be captured automatically without prior input search terms; however, although methods like concgram search can return all instances of all word associations in both forward and backward directions, to tell apart verb–object and/or verb–complement disyllabic combinations from all sorts of word combinations may require prohibitively substantial manual work.

empirical data are certainly a useful complementary resource as well. A closer examination of the 166 SC types demonstrates that basic concepts and categories in everyday life attract more SC uses, and these SCs are repeated by language users (refer to Section 4.1 for further discussion).

A total of 2793 concordance lines of SCs were returned from the 2 million word corpora using a Perl (Practical Extraction and Retrieval Language) script that we wrote for this project, among which 1348 instances are crude SCs (609 forward instances + 739 backward instances) in the written corpus of the LCMC, and 1445 instances of crude SCs (686 forward + 798 backward) in the spoken corpus of the LLSCC. The following is a sample concordance line of the split use of *shui4jiao4*, 'to sleep'.

(6) 睡觉

她倒打起呼噜来了，我倒睡不着了，啊呀，这晚上 *睡*不着*觉*，白天就

休息不好，精力不足呀

shui4jiao4 *Ta1 dao4 da3 qi3 hu1lu lai2-le, wo3 dao4 shui4 bu4zhao2-le, a1ya1, zhe4 wan3shang4* ***shui4*** *bu4zhao2* ***jiao4****, bai2tian1 jiu4 xiu1xi4 bu4hao3, jing1li4 bu4zu2 ya*

sleep she ADV start snore ASP, I ADV can't sleep ASP, INT, this night sleep not RVC sleep, daytime, so rest not good, energy not full PART

sleep She was heard snoring, and I, instead, couldn't sleep this time. You know, 'Since I couldn't go to sleep, I would become very sleepy and couldn't work properly.'

Next we evaluated each instance of a potential SC manually to remove noise in the automatically processed data; however, noise results such as those illustrated in (7) are not rare in the 2793 crude concordance lines.

(7) 见面

只 **见** 区委 书记 **面** 朝 人群， 两膝 跪下， 双手 抱拳， 一连 对

天 作 了 10 个 揖

jian4mian4 *zhi3* ***jian4*** *qu1wei3 shu1ji4* ***mian4*** *chao2 ren2qun2, liang3xi1 gui4xia, shuang1shou3 bao4quan2, yi4lian2 dui4 tian1 zuo4-le 10 ge yi1*

meet ADV see district secretary face toward crowd, two knees kneel down, two hands form fist, consecutively toward sky make ASP 10 CL salutes.

meet The Party Secretary of the District faced 'the crowd, knelt down, two hands holding together, and made 10 salutes to the sky.'

We are looking for *jian4mian4*, 'to meet someone'. In example (7), however, *jian4* (to see) and *mian4* (n. face) are also within the span of 10 token distance, thereby being captured by our algorithm. An English phrasal verb example can better explain the noise sequences in question. For example, if the target phrases are the different forms of "take over", then (8a) and (8b) are what we wish to see, but (8c) and (8d) are phrasal verbs in disguise.

(8) (a) John *took over* his place.
(b) John *took* his place *over*.
(c) John *took over* a year to get through all these.
(d) John *took* the larger one *over* there.

Similar noise cases were removed manually one by one. 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.

The third step of our procedure was to build a database of concordances with distributional information. The distributional information of each sub-genre is also calculated, which will be discussed later. Then we developed an annotation scheme (see Appendix I) that encodes insertion type, direction of separation, semantic type, pragmatic meaning, clause type, discourse function, genre, etc. All the annotation was done by the second author of the paper and double checked by the third author to ensure inter-annotator reliability. Eighty-one tags are used for the annotation scheme. Example (9) is a sample of categories annotated in

Table 1
Distribution of SCs in LLSCC and LCMC.

Corpora	Types	Tokens[a]
LLSCC	108	327
LCMC	104	238
Total	166	565

[a] Throughout the paper, the calculation of SC type and SC token counts the number of SC_H and SC_T pairs, not the insertion words or grammatical categories.

our SC database. It shows the coding of the grammatical categories and highly grammaticalised lexical items separating the SC_H and the SC_T.

(9) <s n="6" genre="WF"> 而年纪较大的人或者已有高血压、血中胆固醇过高、体重过重 或吸烟过量的人，由于他们已有发作心脏病的潜伏危机，如果在运动后就马上<TAG form="JM" direction="F" wordSemantics="0" sentenceSemantics="N" sentenceType="DEC">洗热水澡</TAG>那就是一种危险了。</s>

In this example, <s n="6" genre="WF"> means that the current sentence appears as the 6th sentence in the text sample, and the genre of text is "WF", the text category F (popular lore) of our written corpus, LCMC. Form = "JM" suggests that this instance of SC has an adjectival modification pattern of inserted constituents. Direction = "F" code distinguishes the $SC_H \cdots SC_T$ (F = forward) or $SC_T \cdots SC_H$ (R = reverse) constructions. Negative, neutral or positive lexical and sentence semantic values are also marked up. The SentenceType code shows whether the SC occurs in a declarative (DEC) or interrogative sentence (QUE) (see Appendix I for a full explanation of the annotation scheme).

Once the data had been thus annotated, we carried out the quantitative and qualitative analyses.

4. Results, findings and discussions

4.1. General distribution

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

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), *gao4zhuang4* (9), *shou4zui4* (8), *qian4zhai4* (6), etc. The top SCs in the LCMC are *du2shu1* (14), *hong2lian3* (10), *wo4shou3* (10), *xi3zao3* (10), *di1tou2* (9), *jian4mian4* (9), *shui4jiao4* (9), *bang1mang2* (8), *feng4ming4* (8), *fan4zui4* (5), etc.

The lexical meaning of the most typical SCs seems to centre on common bodily sensations (e.g. to blush, to get angry or shy), basic everyday human activities (e.g. to sleep, to help, to meet, to shake hands, to take a shower), which actually have acquired a metaphorised meaning after repeated use over time (e.g. "face red" now means "to get 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 0.001 (see Table 2). This suggests that SCs are significantly more frequent in spoken Chinese than in written Chinese. The high 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.

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

	LLSCC (1,002,151)	LCMC (1,006,731)	LL
SC types	108	104	0.09
SC tokens	327	238	14.49

Table 3
Top 20 SCs in LLSCC and LCMC.

Freq.	SCs	Freq.	SCs
33	*bang1mang2*, help busy, 'to help'	10	*shou4zui4*, receive penalty, 'to have a hard time'
31	*du2shu1*, read book, 'to read; to learn or study'	10	*di1tou2*, lower head, 'to give up, to succumb to'
27	*jian4mian4*, see face, 'to meet'	9	*gao4zhuang4*, report testimonial, 'to complain to the higher authorities'
23	*shui4jiao4*, sleep sleep, 'to sleep'	8	*zhang1zui3*, open mouth, 'to talk, to ask for'
21	*nian4shu1*, read (aloud) book, 'to learn, to study'	8	*ting1ke4*, hear class, 'to attend classes'
15	*lai2xin4*, come letter, 'to hear from'	8	*feng4ming4*, hold order, 'to act under orders'
13	*xi3zao3*, wash bath, 'to bathe'	8	*bai4shi1*, salute master, 'to acknowledge sb. as one's master'
13	*hong2lian3*, red face, 'to feel unhappy about sb.'	7	*qian4zhai4*, owe debt, 'to owe sb. (money)'
12	*hui2xin4*, back letter, 'to write back to sb.'	7	*ju1gong1*, bow bow, 'to bow sb.'
11	*wo4shou3*, shake hand, 'to shake hands'	7	*fan4zui4*, commit crime, 'to commit crime'

104). A further analysis of the actual SC types used in the LLSCC and the LCMC shows that the two datasets share the top SCs, only with slightly different ranking orders. This indicates that the spoken tendency of SCs is basically the result of more frequently repeated uses in spoken Chinese.

This register difference regarding SC usage is of interest; however, of even greater interest is the breakdown of the frequency distribution of SCs in the genres covered in the spoken and written corpora. Tables 4 and 5 show the distribution of SCs across spoken and written genres respectively (Figs. 2 and 3).

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. This in a sense might shed light on the grammaticalisation of SCs from speech to writing. Diachronic data might show that change in spoken language heralds the emergence of new forms and usage in the written language. Even if this grammaticalisation were nothing but a linguist's hunch, in the least an immediate explanation to the conceivable distribution of SCs in different text types can be attributed to the information-versus-involvement focus of spoken and written genres (cf. Biber, 1988, pp. 104–108). Both TV and movie scripts (performances of TV plays, soap operas and movies) 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 (mainly televised university debating competitions) aim to convey a considerable amount of information within limited time.

What is then the source of the overall genre variation remains at issue. In the absence of comparable historical data, all that we can investigate is the synchronic linguistic behaviour of SCs *per se*, the general lexical

Table 4
SC distribution in sub-genres of LLSCC.

	Text category	Raw freq.	Normalised freq. (per min)
S01	Direct conversation (60,806 words)	6	98.7
S02	Telephone conversation (295,026 words)	62	210.2
S03	TV and movie scripts (80,446 words)	51	***634.0***
S04	TV talk show (118,588 words)	34	286.7
S05	Oral narrative (102,262 words)	31	303.1
S06	Edited oral narrative (267,114 words)	140	524.1
S07	Formal debates (77,909 words)	3	***38.5***

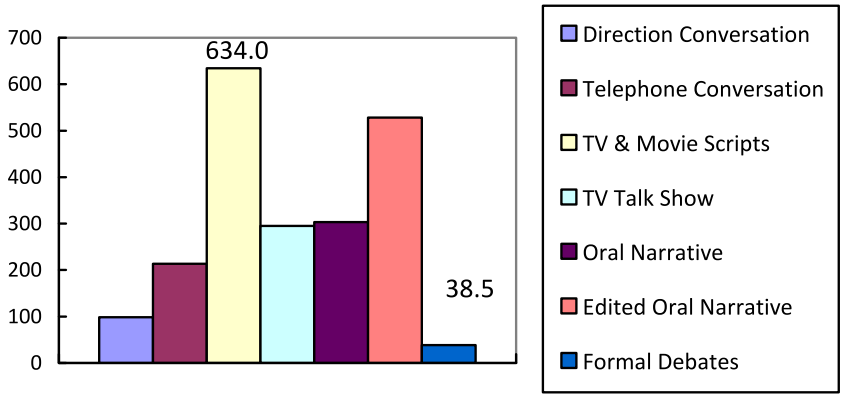

Fig. 2. SC distribution in sub-genres of LLSCC.

Table 5
SC distribution in sub-genres of LCMC.

	Text category	Raw freq.	Normalised freq. (per min)
W01	News reportage (88,333)	21	237.7
W02	News review (53,902)	3	55.7
W03	Religion (34,401)	4	116.3
W04	Trade/skill/hobby (76,558)	9	117.6
W05	Popular lore (88,392)	25	282.8
W06	Biography (154,646)	49	316.9
W07	Report/official document (60,795)	4	65.8
W08	Academic prose (160,515)	4	***24.9***
W09	General fiction (58,432)	26	445.0
W10	Mystery and detective story (48,422)	23	475.0
W11	Science fiction (12,356)	2	161.9
W12	Martial arts fiction (58,421)	21	359.5
W13	Romantic fiction (58,367)	31	531.1
W14	Humour (18,495)	16	***865.1***

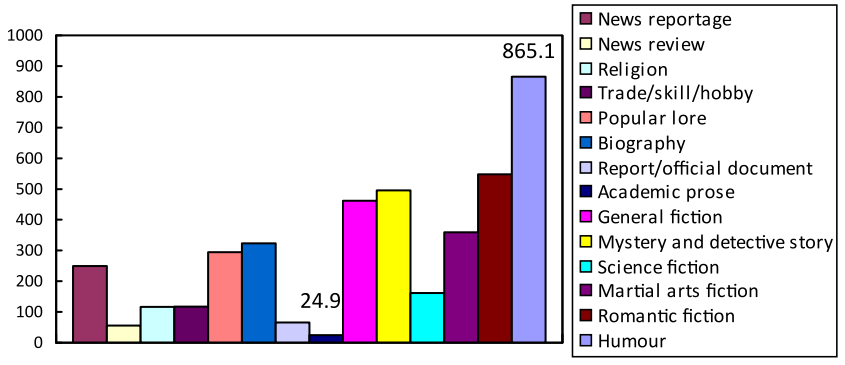

Fig. 3. SC distribution in sub-genres of LCMC.

semantics of the construction and/or components of them, and more intriguingly, what may separate the component morphemes of SCs from each other. It is to this that we now turn.

4.2. Lexical and grammatical patterning of insertions

The elements between SC_Hs and SC_Ts are referred to as insertions in this paper. Their length varies. The corpus analysis will provide the relevant calculation of the span of insertions. Intuitively it seems to be no

Table 6
Type 1a "SC_H ASP SC_T" constructions in LLSCC and LCMC.

	SC types (%)	SC tokens (%)	SC TTR
SC_H-*le* SC_T	42 (25%)	74 (13%)	0.58
SC_H-*guo* SC_T	15 (9%)	22 (4%)	0.68
SC_H-*zhe* SC_T	12 (7%)	35 (6%)	0.34
Total	69 (42%)	131 (23%)	0.53

longer than 10 tokens.[5] 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. The lexical and grammatical patterning of SC insertions may be viewed as practically internal "collocations" (Firth, 1951/1957, p. 194) and "colligations" (Firth, 1957/1968, p. 181) if we see SC_Hs and SC_Ts as individual lexical items.

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 6.

Of the 575 instances of SCs in the two corpora 13% (74) involved "SC_H-*le* SC_T" of which there are 42 different types. We hereby call the three-morpheme pattern "SC_H ASP SC_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'
tie3-le *xin1* (6), iron ASP heart, 'to be determined'
shou4-le *zui4* (4), receive ASP penalty, 'to have had a hard time'
hong2-zhe *lian3* (10), red ASP face, 'to be feeling unhappy about sb.'
di1-zhe *tou2* (10), lower ASP head, 'to be lowering one head'
jian4-guo *mian4* (4), see ASP face, 'to have met'
shui4-guo *jiao4* (2), sleep ASP sleep, 'to have slept'
nian4-guo *shu1* (2), read ASP book, 'to have received schooling'

Among the most common SC_H ASP SC_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. For instance, in addition to its basic lexical meaning of sleeping, *shui4*-guo *jiao4* can be particularly negative in its connotation, implying a woman sleeping with another man, which is a disgraceful act. If the original perfective meaning of "to have slept" has to be expressed, it must be preceded by a negator (*mei2*, *mei2you3*, or *bu4*, as in *hao3 ji3tian1 mei2 shui4*-guo *jiao4*, 'to have not slept for many days'). In other words, the literal meaning of *shui4*-guo *jiao4* is possible, but the most likely meaning is the idiomatic one. The following two examples in (11) and (12) are from spoken and written narratives respectively.

(11) 她 问 我 那时 和 几 个 男孩子 *睡 过 觉*。
Ta1 wen4 wo3 na4shi2 he2 ji3ge nan2hai2zi ***shui4-guo jiao4***.
she ask I that time with how many CL boys ***sleep ASP sleep***
'She asked how many boys with whom I had ever slept at that time.'
(from LLSCC)

(12) 自从 她 15 岁 起，每 一个 和 她 *睡 过 觉* 的 男人 都 是 既 年轻 又 英俊。
Zi4cong2 ta1 15 sui4 qi3, mei3 yi1ge4 he2 ta1 ***shui4-guo jiao4*** *de nan2ren2, dou1shi4 ji4 nian2qing1 you4 ying1jun4*.

[5] Here tokens are monosyllabic and compound words in general and in actual analysis refer to the words identified by our automatic tokeniser.

since she 15 years old ASP, every CL with her ***sleep ASP sleep*** DE men, all both young and handsome.
'All men with whom she slept since she was 15 were both young and handsome.'
(from LCMC)

This "to sleep" SC is reminiscent of the English word "born", the base form of which "to bear" is much less readily used than the form "born".

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.

(13) a. 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…'
b. 她 给 我 已经 *来 过 一 封 信* 了 (from LLSCC)
Ta1 gei3 wo3 yi3jing1 **lai2-guo yi1 feng1 xin4** *le*
she give me already ***come ASP a CL letter*** ASP
'She has ***sent me a letter already***.'
c. 向 大家 深深地 *鞠 了 一 躬* (from LCMC)
Xiang4 da4jia1 shen1shen1de **ju1-le yi1 gong1**
to everybody deeply ***bow ASP a bow***
'***To have taken a deep bow*** to everybody'

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_H-*le* SC_T, a three-morpheme verbal construction with an actual aspect marker in the middle. The pattern in question seems to be in line with the probabilistic curve of Chinese words proposed in the seminal work by Zipf (1935, p. 26, p. 45): a small number of shorter words account for the majority of the entire Chinese lexicon. The length of Chinese words is in reverse proportion to their relative token frequency (see Xiao et al., 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. There does not seem to be any theory *a priori* which could well explain this except that a usage-based emergent view (Bybee, 2007; Bybee and Hopper, 2001; Hopper, 1987; Tao, 2003) is able to explain the probabilistic "orderliness" (Zipf, 1935, p. 48).

In addition to aspect markers, quite a few other elements can be attached to the SC_Hs. 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 instances in LCMC/LLSCC), *xia4*, 'down, downward' (3/26), *shang4*, 'up, upward' (3/26), *hao3*, '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).

The 0.77 type/token ratio (TTR) in Table 8 suggests that the SC_H RVC SC_T type is quite productive, much more flexible than the SC_H ASP SC_T type (TTR, 0.53) or the SC_H (?) ASP (?) SC_T type (TTR, 0.37). However,

Table 7
Type 1b "SC_H (?) ASP (?) SC_T" constructions in LLSCC and LCMC.

	SC types (%)	SC tokens (%)	SC TTR
SC_H (?)[a] ASP (?) SC_T	91 (55%)	244 (43%)	0.37

[a] Here the (?) in "SC_H (?) ASP (?) SC_T" indicates that this slot may be filled either by other elements or left blank. This notation is observed in later analyses as well.

Table 8
Type 2a "SC_H RVC SC_T" insertions.

	SC type (%)	SC token (%)	SC TTR
SC_H RVC SC_T	20 (12%)	26 (5%)	0.77

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. Like the expandable cases of aspect-marker insertions, RVC insertions have their variants as well, which are shown in Table 9. Among all instances of the SC_H (?) RVC (?) SC_T type, there are some negated SC constructions (i.e. SC_H BU RVC (?) SC_T, see example 14), which occur 21 times (4% of the total SCs).

(14) a. 念 不 好 书 (from LLSCC)
Nian4 bu4 hao3 shu1
read NEG well book
'Does not do well in school'
b. 帮 不 上 忙 (from LLSCC)
Bang1 bu2 shang4 mang2
help NEG RVC help
'Cannot be helpful'
c. 睡 不 好 觉 (from LCMC)
Shui4 bu4 hao3 jiao4
sleep NEG well sleep
'Cannot sleep well'
d. 张 圆 小 嘴 (from LCMC)
Zhang1 yuan2 xiao3 zui3
open round little mouth
'With his/her little mouth rounded'

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 SC_Ts, namely the nominal/complement components. Most typical items in this category are quantificational expressions, classifiers, and various forms of pre-modifiers. The following section will describe the behaviours of the SC_T satellites.

4.2.2. Nominal/complement satellites of SC constructions

4.2.2.1. 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. *ji3*, 'several'; *liang3*, 'a couple of') or exact numbers (e.g. *yi1*, 'a, one'; *san3*, 'three'). In Chinese, especially in everyday language, *liang3*, 'two', can mean two to three, or "a couple of", when it is unstressed. Similarly, *yi1* can refer to an exact number, one; or very often it does not have any reference to the actual amount, as in *chi1 yi1 fan4*, 'have a meal' in Beijing Mandarin. As the transliteration shows, this *yi1* resembles, to some degree, the indefinite article in English. Its meaning of

Table 9
Type 2b "SC_H (?) RVC (?) SC_T" insertions.

	SC types (%)	SC tokens (%)	SC TTR
SC_H (?) RVC (?) SC_T	20 (12%)	66 (12%)	0.30

amount is very weak, and its grammatical function stands out. However, *yi1* in this usage is not recognised as an article, given that Chinese does not actually have such a grammatical category. This unstressed *yi1*, as well as *yi1ge* in some cases, does have its diminutive or mitigating discourse-pragmatic function (Feng, 2002; Biq, 2004), which will be discussed in Section 4.4. Moreover, demonstratives like *zhe4*, 'this' (16 times, five occurrences of *zhe4* used alone, other variants and related forms of *zhe4* like *zhe(i)4ge*, *na4xie1*, *zhe4yang4*, etc. are not distinguished here) and *na4(ge)*, 'that' (8 times) are likely to occur in the same slot prior to the classifier and the nominal element.

4.2.2.2. 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 *ge* which occurs 40 times and accounts for 34% of all the quantifiers among the split SCs. The versatile classifier *ge*, which can be generally understood as "piece" in English, is not easily translatable into a non-classifier language such as English. The highly grammaticalised *ge* has undergone considerable semantic bleaching. It has, however, acquired its pragmatic meaning. A single *ge* can be used in all our 166 SCs, forming a three-morpheme construction like $SC_H + ge + SC_T$, although they are not actually all materialised as such. We have found 18 occurrences of $SC_H + ge + SC_T$s in our data, e.g. *bang1 ge mang2* (4), *xi3 ge zao3* (2), *qian1 ge zi4* (2) and *jing4 ge li3* (2). They are most likely to be found in spoken Chinese (e.g. all four instances of *bang1 ge mang2*, two instances of *qian1 ge zi4* and *jing4 ge li3* occur in LLSCC), or the fictional genres of written Chinese (e.g. 2 times of *xi3 ge zao3* in romantic fiction and humorous texts).

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. Such modifying elements, in our data, also include personal possessive pronouns, which are regarded as content words, instead of function or determiner type of items in Chinese grammar. Following the same tradition, personal possessive pronouns will be discussed in the modifier category of insertions.

4.2.2.3. 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. *shen2me*, 'what', etc. 26 times, 5% of all SCs), and also combinations of these elements as illustrated in the following examples.

(15) Adjectival modifier
bang1-le na4ge ren2 DA4 mang2, help ASP that person BIG help, 'did that person a big favour'

(15) Nominal modifier
xi3 RE4SHUI3 zao3, wash HOT WATER shower, 'had a hot water shower'

(16) Possessive personal pronoun
ting1 WO3DE ke4, listen MY lesson, 'attend my class'

(17) Question word[6]
chui1 SHEN2ME niu2, blow WHAT cow, 'boast about something'

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 *de* is used between the modifier and the SC_T. In authentic data, it is not surprising to see repetition, false starts or swear words in insertions. These residual categories will not be discussed here.

[6] Question words, especially *shen2me*, within SCs carry a negative meaning.

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.

Paradigmatically, the 166 instances of the "$SC_H \ldots SC_T$" 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. Fig. 4 depicts fully developed discontinuous usage, but obviously in reality SCs are realised in various forms of different length and different combinations. In the figure, the SC_H is the verbal head and SC_T the nominal/complement element of an SC. The "$SC_H \ldots SC_T$" 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), various forms of pre-modifiers. Alternatively, the insertion of a discontinuous SC can be a combination of two or more elements listed in Fig. 4. Fig. 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. Although the major categories are all represented in the diagram, the weight of the most frequent usage of SCs tips towards the upper left-hand portion of the diagram, given that the three-morpheme aspect-marker insertion type takes up 42% of all SCs. In other words, discontinuous SCs do not have to be very long and complex. 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.

The diagram gives a skeletal but explanatory description of the lexical and grammatical patterning of SCs based on the 2 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. Unlike in the case of German separable verbs as in the example, *wann fangen Sie an*, 'when do you start', in which the position of the affix can go before or after the verbal head, Chinese SCs follow a relatively stable order, SC_H before SC_T.

$$\mathbf{SC_H} + \text{NEG} + \text{ASP/RVC} + \text{MC} + \text{CL} + \text{MOD} + \mathbf{SC_T}$$

Fig. 4. A syntagmatic pattern of typical SCs.

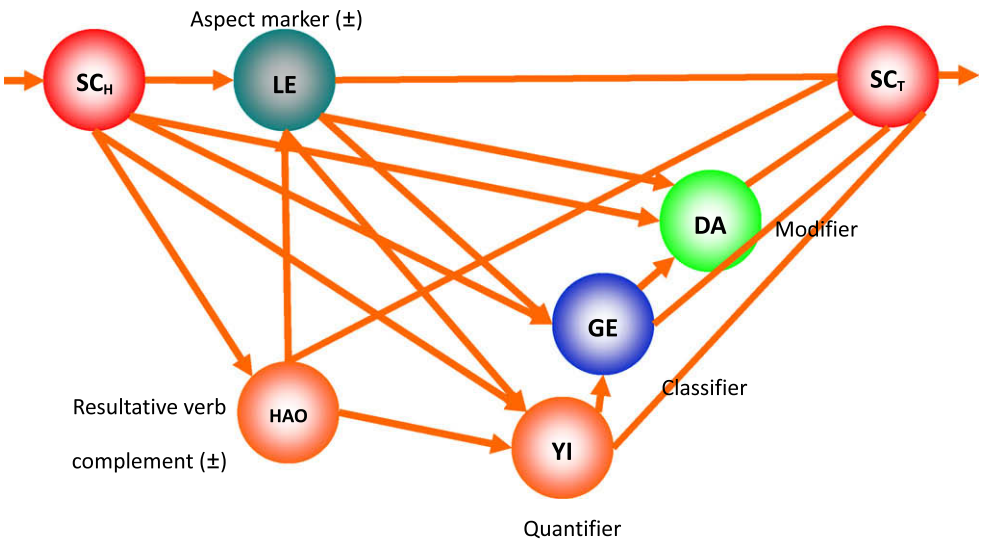

Fig. 5. Typical lexical and grammatical patterning of discontinuous SCs. (The diagram is inspired by Mindt (1995, 2000, 2002), in which he plotted the interactive relations within verb and auxiliary verb systems in English.)

We are not suggesting that the German phenomenon is the same as the Chinese. We are fully aware that the German phenomenon is not local and may involve complex and recursive structures while the Chinese phenomenon is local and restricted in the way we have outlined.

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, namely aspectual 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', e.g. jian3zhi2 shui4 bu4 liao3 jiao4, simply sleep not ASP sleep, 'can hardly fall asleep'. 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_Hs and SC_Ts 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_Hs and SC_Ts. 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. Our decision of negative sentences was based on the presence of both SC internal and SC external negators. When two negators were found, one inside the SC_H and SC_T template, the other in the clause outside the SC_H and SC_T, the sentence semantics was read as positive. The respective frequencies of the six types of sentences are presented in Table 10. 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_Hs and SC_Ts apart.

4.4. Discourse-pragmatic properties of SCs and their insertions

The high propensity of SCs in speech has already shed some light on the discourse preference of SCs. To make sure about the correlation of discontinuous SCs with discourse modes, 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 2 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. The sub-genres of spoken data do not present a consistent more-formal–more-frequent distribution of SCs, either. However, in the written data, news review (22 times, or 408 times per million words, 1.6% of the written corpus) and academic prose (87 times, or 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, 2788 times per million words, 11% of the written corpus), and humorous texts (61 times, 3298 times per million words, 13% of the

Table 10
Sentential context of SCs.

Frequencies (%)	Sentence type
455 (81%)	Declarative
32 (6%)	Negative
27 (5%)	Exclamatory
26 (5%)	Interrogative
22 (4%)	Rhetorical
3 (1%)	Imperative

written corpus) are two text types which contain most combined uses. Therefore, both discontinuous and continuous use of SCs point to the spoken genre preference of SC usage, which is particularly true of split usage of SCs. Likewise, we noted in our data the strong collocation between SC_Hs/SC_Ts and aspect markers (e.g. the actual marker *-le* and experiential marker *-guo*), which are commonly used in describing past events. According to Biber (1988, pp. 108–109), the frequent use of perfect aspect verbs is characteristic of narrative discourse, where a sequential account of past events is presented. In addition to the temporal ordering of narrative discourse, post-verbal resultative complements assist to modify the intensity or degree of an activity so as to be more specific or particularised. This elaborated modification of verbs may have to do with the speaker's involvement in the discourse (cf. Biber, 1988, p. 104; Ljung, 2002, p. 181).

Aspect markers, as well as resultative verb complements which too convey aspectual meaning, are found in over half of the total occurrences of SCs. Considering their strong association with the verbal heads, there is good reason to view SCs as expressive devices in producing elaborated real life stories. However, what does the great variety of satellite grammatical elements preceding the SC_Ts have to offer to the discourse? The two parts of the construction, in the first place, combine with each other to form a coherent unit of meaning. In the meantime, they function independently with their respective associates. The frequently used lexico-grammatical categories such as quantifiers, classifiers and different modifiers typically precede SC_Ts. Also, 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_Ts. A classifier prior to a noun is very often obligatory in Chinese. Once it is dropped, it implies a genre-specific usage, which is more often found in Beijing Mandarin (e.g. *chi1 yi fan4*, 'to have a meal' instead of *chi1 ge fan4* or *chi1 yi1 ge fan4*). The high frequency *ge* is also textually or pragmatically primed. According to Biq (2004, p. 1663, see also Biq, 2002, 2007 for more discussion), 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 "V1 (*yi*) *ge* V2 construction provides a 'buffer zone' in the on-line production situation". The classifier (*yi*) *ge*, together with other adjectival and nominal modifiers, seems to, in most cases, mitigate, soften, or qualify the properties of the ensuing nominal elements. This spoken discourse preference can be deduced from our comparison of spoken and written Chinese data. The insertions as illustrated in Fig. 5 help to express a more personal stance than modification alone. The most frequent SC type in our data, *bang1mang2* (help busy, 'to help'), is such a case in regard to the different discourse-pragmatic functions of split SCs. For instance, in the construction *bang1 ge mang2* (help CL busy, 'to help') the inserted classifier *ge* mitigates the tone of a request for a favour; however, in the case like *bang1*-le *yi1 ge da4 mang2* (help ASP a CL busy, 'to offer a big helping hand') the inserted part serves as a comment on the big favour to the beneficiary.

5. Discussion and conclusion

It is clear, from the examination of the lexical, grammatical and discoursal properties of SCs, that SCs in Chinese straddle the morphology/syntax divide and also the structure vs. discourse one. Based on a quantitative and qualitative 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 their morpho-syntactic status.

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). This gestalt lexico-semantic status has been the justification for the sinogram-based approach to SCs (see Section 2.3.2). Moreover, the syllable morpheme correspondence makes it possible to study both grammatical and phonological aspects, as suggested in Dixon and Aikhenvald (2002), of morphology at once. Hence, 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. The structural account of SCs, however, maintains that to be a word (compound in this case), 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_Ts are always to be found after certain SC_H verbal heads, either immediately adjacent to them or in close proximity.

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. This will provide a more nuanced perspective of SCs than has been available to date. 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 given the Foot Formation Rule (FFR), with disyllabicity as its predominant form. Thus the combined uses of SCs are compounds. Feng also recognises that compounds sometimes allow the insertion of modifiers (Feng, 2002, p. 96), but he has no clear basis for assigning wordhood to the separated SCs.

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 2 million words of spoken and written data, the following typology of the manifestations of SCs:

(a) Combined uses of SC_Hs and SC_Ts will be argued to be compounds on the basis of the FFR;
(b) SC_Hs and SC_Ts separated by one single aspect marker will be considered as compounds according to the so-called Mending Device of Trisyllabic Foot (Feng, 2002, p. 112), given that an aspect marker may be viewed as a counterpart of an inflectional suffix following a verb in a morphologically inflectional language; while
(c) SC_Hs and SC_Ts separated by other grammatical categories like numerals, quantifiers, adjectival modifiers, etc. will be analysed as phrases.

Our quantitative data explicate that over half (see Table 7) 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.

5.1. 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 SC_Ts, represented typically by a noun or complement, as least possibly compounds.

5.2. 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 SC_Hs and SC_Ts are disyllabic compounds;
(b) SCs in which the SC_H and the SC_T are separated by one single morpheme under the Trisyllabic Foot Rule are possible compounds; while
(c) SC_Hs and SC_Ts 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. For instance, the three-syllable *chi1 ge fan4*, 'to have a meal', is on the far left of the phonological cline, but on the far right of the structural cline, because *ge* is a nominal pre-modifier rather than an aspect marker attached to the preceding verb head, and thus it is not a good example of an SC, though it is a trisyllabic construction. So when there is a mismatch between a grammatical word and a phonological word, we would give priority to the host dependency criteria.

This study has arrived at some interesting generalisations about SCs on the basis of their textual distribution, the general lexical semantics of SC_Hs and SC_Ts, and the interaction between the two. From the spoken

and written Chinese data, a relatively explicit cline of typicality of SC uses has emerged which may serve as a source for future comparisons both of a diachronic nature (if relevant diachronic data prove to be available) or dialectal ones, relating to different varieties of Chinese. Since our aim was essentially a descriptive one, we have not pursued any specific theoretical account of SCs. The continuum of SCs that we have suggested does, however, lay down the empirical foundation for a theoretical analysis.

Acknowledgements

We are greatly obliged to ESRC grant (RES-000-22-2286), without which this study could not have been possible. This study is also supported by the National Research Centre for Foreign Language Education (MOE Key Research Institute of Humanities and Social Sciences at Universities), Beijing Foreign Studies University and the Chinese MOE research grant (Ref.: 08JC740002). We are particularly grateful to the participants at the 41st Annual Meeting of the *Societas Linguistica Europaea*, as well as the audience at the Verb Typologies Revisited conference, for their critical comments. We would thank two anonymous reviewers for their very constructive comments on the manuscript.

Appendix I. Annotation scheme

AUX RM	(AUX, e.g. modal verb)
BU	不
BU AUX	
CL	classifier
CL JM	
COMPLEX	very complicated structure
DE	的 as a predicative marker
DE1 JM	
DE3	得 (的)
DIAN	点, 点儿 indicating quantity or amount
DIAN JM	
FOC	focus marker (连, 都, 就是, etc.)
FOC BU	
FOC BU AUX	
FOC BU OBJ	
FOC PI	
FOC ZAI	
GUO	过 as the experiential aspect marker
GUO JM	
GUO MC	
JM	adjectival modifier
JM MC	
JM SM	
LE	了 as the actual aspect marker
LE CL	
LE CL JM	
LE JM	
LE MC	
LE OBJ JM	
LE OBJ MC	
LE SW	
LE YI	

LE YI CL	
MC	quantifier (may include a classifier; but excludes 一)
MEI	没，没有
OBJ	object
OBJ SM	
OBJ YI	
OBJ YI CL	
PI	potential infix structure (-得-，-不-)
PI JM	
PI MC	
PI SM	
QILAI1	起来 as the inceptive aspect marker
QILAI2	
QILAI2	起来 (result)
RM	adverbial modifier
RVC	resultative verb complement
RVC LE	
RVC LE JM	
RVC MC	
RVC YI	
SM	什么，啥，何, etc.
SM JM	
SUB RM	
SW	swear word
You	aspect marker
YI	一
YI CL	
YI2	一 (一旦)
ZHE	着 as the durative aspect marker
Direction	
F	forward (normal splittable compounds)
B	backward (reverse splittable compounds)
WordSemantics	
N	negative
0	neutral
P	positive
SentenceSemantics	
N	
0	
P	
SentenceType	
DEC	declarative
NEG EXC	negative + exclamation
RHE	rhetoric question
IMP	imperative
NEG	negative
EXC	exclamation
QUE	question
IMP NEG	imperative + negative

Appendix II. 166 SC types found in LLSCC and LCMC

帮忙	吵架	享福	旷课
读书	铁心	梳头	看中
见面	跳舞	平反	敬礼
睡觉	迈步	聊天	怀孕
念书	练功	静心	画像
来信	吸烟	换钱	关门
洗澡	问话	丢人	鼓掌
红脸	举例	种田	搞鬼
回信	操心	中邪	负债
握手	毕业	征兵	翻身
受罪	站住	摇头	发音
低头	抬头	许愿	辞职
告状	签字	现眼	成事
张嘴	教书	吓人	唱戏
听课	兼职	忘掉	猜透
奉命	嫁人	送信	搬家
拜师	划清	伤心	摆手
欠债	点头	容人	碍事
鞠躬	吹牛	瞧见	组团
犯罪	炸锅	拼命	撞车
动心	游泳	录音	装蒜
当事	咬牙	理发	转向
住嘴	输液	留影	改线
争气	使劲	亮相	封门
找主	生事	炼油	费神
展开	升旗	劳神	堵嘴
遭罪	伸腰	亏本	跌交
圆梦	上瘾	坑人	低头

要命	任职	考取	担心
演戏	让座	揩油	带信
言声	劝酒	开张	闯祸
悬心	牵线	进站	扯皮
叙旧	谱曲	尽职	测字
泄气	撇嘴	加油	变样
误事	碰头	记事	抱团
秃顶	跑步	混饭	拌嘴
透风	拍照	化妆	办学
通信	努嘴	合影	把关
挑刺	募捐	过瘾	爱美
题词	迷路	共事	挨整
套汇	没辙	弓腰	
甩手	埋头	赶路	

References

Biber, Douglas, 1988. Variation across Speech and Writing. Cambridge University Press, Cambridge.

Biq, Yung-O., 2002. Classifier and construction: the interaction of grammatical categories and cognitive strategies. Language and Linguistics 3 (3), 521–542.

Biq, Yung-O., 2004. Construction, reanalysis, and stance: 'V yi ge N' and variations in Mandarin Chinese. Journal of Pragmatics 36 (9), 1655–1672.

Biq, Yung-O., 2007. Buding liangci ciyi yu goushi de hudong (The interaction of classifiers and construction). Zhongguo Yuwen (Chinese Language) (6), 507–515.

Branner, David, 2003. On early Chinese morphology and its intellectual history. Journal of the Royal Asiatic Society 13 (1), 45–76.

Bybee, Joan, Hopper, Paul (Eds.), 2001. Frequency and the Emergence of Linguistic Structure. John Benjamins, Amsterdam.

Bybee, Joan, 2007. Frequency of Use and the Organization of Language. Oxford University Press, Oxford.

Ceccagno, Antonella, Basciano, Bianca, 2007. Compound headedness in Chinese: an analysis of neologisms. Morphology 17 (2), 207–231.

Chao, Yuan Ren, 1968. A Grammar of Spoken Chinese. University of California Press, Berkeley.

Cheng, Winnie, Greaves, Chris, Warren, Martin, 2006. From N-gram to skipgram to concgram. International Journal of Corpus Linguistics 11 (4), 411–433.

Dixon, Richard, Aikhenvald, Alexandra (Eds.), 2002. Word: A Cross-linguistic Typology. Cambridge University Press, Cambridge.

Feng, Shengli, 2001. Cong yunlu kan hanyu ciyu fenliu zhi dajie (Prosodically-determined distinction between word and phrase in Chinese). Zhongguo Yuwen (Chinese Language) (1), 27–37.

Feng, Shengli, 2002. The Prosodic Syntax of Chinese. Lincom Europa, Muenchen.

Firth, John, 1951/1957. Modes of Meanings. Reprinted in Papers in Linguistics 1934–1951. Oxford University Press, London, pp. 190–215.

Firth, John, 1957/1968. A synopsis of linguistic theory, 1930–55. Studies in Linguistic Analysis (Special Volume of the Philological Society). Reprinted in Frank Palmer. Selected Papers of J.R. Firth 1952–59. Longmans, Green and Co., Ltd., London and Harlow, pp. 168–205.

Hopper, Paul, 1987. Emergent grammar. Berkeley Linguistic Society 13, 139–157.

Hu, Yushu, Fan, Xiao, 1996. Dongci Yanjiu Zongshu (A Survey of Studies on Verbs). Shanxi Gaoxiao Lianhe Chubanshe (Shanxi United Press of Universities), Taiyuan.

Huang, James C.-T., 1984. Phrase structure, lexical integrity, and Chinese compounds. Journal of the Chinese Language Teachers Association 19, 53–78.
Li, Charles, Thompson, Sandra, 1981. Mandarin Chinese—A Functional Reference Grammar. University of California Press, Berkeley.
Ljung, Magnus, 2002. What vocabulary tells us about genre differences: a study of lexis in five newspaper genres. In: Breivik, Leiv, Hasselgren, Angela (Eds.), From the Colt's Mouth... and Others'. Rodopi, Amsterdam, pp. 181–196.
Lü, Shuxiang, 1961. Xiandai hanyu dan shuang yinjie wenti chutan (An investigation of the issue of monosyllabicity and disyllabicity in modern Chinese). Zhongguo Yuwen (Chinese Language) (1), 10–22.
Lü, Shuxiang, 1979. Hanyu yufa fenxi wenti (Problems in grammatical analysis in Chinese). Reprinted in Shuxiang, Lü. (Ed.), 1984. Hanyu Yufa Lunwenji (Anthology of Chinese Grammar). The Commercial Press, Beijing.
Lu, Zhiwei, 1957. The Word-formation of Chinese Language. The Scientific Press, Beijing.
Mathews, Peter, 1991. Morphology, second ed. Cambridge University Press, Cambridge.
McCarthy, John, Prince, Alan, 1993. Prosodic morphology I: Constraint interaction and satisfaction. MS. University of Massachusetts, Amherst and Rutgers University.
McCarthy, John, Prince, Alan, 1995. Prosodic morphology. In: Goldsmith, J. (Ed.), Handbook of Phonology. Blackwell, Oxford, pp. 318–366.
McEnery, Anthony, Xiao, Richard Zhonghua, Tono, Yukio, 2006. Corpus-Based Language Studies: An Advanced Resource Book. Routledge, London.
McEnery, Anthony, Xiao, Richard Zhonghua, Mo, Lili, 2003. Aspect marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for contrastive language study. Literary and Linguistic Computing 18 (4), 361–378.
Mindt, Dieter, 1995. An Empirical Grammar of the English Verb: Modal Verbs. Cornelsen, Berlin.
Mindt, Dieter, 2000. An Empirical Grammar of the English Verb System. Cornelsen, Berlin.
Mindt, Dieter, 2002. A corpus-based grammar for ELT. In: Kettleman, Bernard, Marko, Georg (Eds.), Teaching and Learning by Doing Corpus Analysis. Rodopi, Amsterdam, pp. 91–104.
Packard, Jerome, 2003. The Morphology of Chinese: A Linguistic and Cognitive Approach. Cambridge University Press, Cambridge.
Pan, Wenguo, 2006. Zibenwei lilun de zhexue sikao (A philosophical investigation into the sinogram-based theory). Yuyan Jiaoxue yu Yanjiu (Language Teaching and Research) (3), 36–45.
Ren, Haibo, Wang, Gang, 2005. Jiyu yuliaoku de xiandai hanyu lihe ci xingshi fenxi (The analysis of splittable compounds in modern Chinese based on the large-scale corpus). Yuyan Kexue (Linguistic Sciences) 4 (6), 75–87.
Shi, Maozhi, 1999. Shubin fuheci de yufa tedian (Grammatical properties of predicate-object type compounds). Yuyan Jiaoxue yu Yanjiu (Language Teaching and Research) (1), 123–134.
Tao, Hongyin, 2003. Toward an emergent view of lexical semantics. Language and Linguistics 4 (4), 837–856.
Wang, William S.-Y., 1973. Chinese language. Scientific American 228, 50–60.
Wang, Li, 1989. Hanyu Yufa Shi (A History of Chinese Grammar). The Commercial Press, Beijing.
Wang, Chunxia, 2001. Jiyu Yuliaoku de Lihe Ci Yanjiu (A Corpus-based Study of Splittable Compounds). M.A. dissertation, Beijing Language and Culture University.
Xiao, Richard Zhonghua, McEnery, Anthony, 2004. Aspect in Mandarin Chinese: A Corpus-based Studies. John Benjamins, Amsterdam.
Xiao, Richard Zhonghua, Rayson, Paul, McEnery, Anthony, 2009. A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners. Routledge, London.
Xu, Tongqiang, 1997. Yuyan Lun (On Language). North China Normal University Press, Changchun.
Yang, Qianhui (Ed.), 1995. Xiandai Hanyu Lihe Ci Yongfa Cidian (A Dictionary of Splittable Compound Usage in Modern Chinese). Beijing Normal University Press, Beijing.
Yang, Xipeng, 2003. Hanyu Yusu Lun (A Theoretical Account of Chinese Morphemes). Nanjing University Press, Nanjing.
Yu, So Sum, 2003. Discontinuous Verb-object Compounds in Cantonese and Mandarin. Unpublished MA dissertation. University of Hong Kong, Hong Kong.
Zhang, Hesheng, 2007. Ye tan duiwai hanyu cihui jiaoxue de benwei zhi zheng (A revisit to the minimal units in Chinese as a Foreign Language). Yuyan Wenzi Yingyong (Applied Linguistics) (4), 2–5.
Zhou, Shangzhi, 2006. Hanyu Lihe Ci Yanjiu: Hanyu Yusu, Ci, Duanyu de Teshuxing (A Study on Chinese Splittable Compounds: The Peculiarities of Chinese Morphemes, Words, and Phrases). Shanghai Foreign Language Education Press, Shanghai.
Zhu, Dexi, 1982. Yufa Jiangyi (Lecture Notes on Grammar). The Commercial Press, Beijing.
Zhu, Kunlin, 2006. Xiandai Hanyu Cidian zhong de Lihe Ci Yanjiu (A study of Splittable Compounds in Modern Chinese Dictionary). Jilin Sheng Jiaoyu Xueyuan Xuebao (Journal of Educational Institute of Jilin Province) 22 (2), 29–30.
Zipf, George, 1935. The Psycho-Biology of Language. The MIT Press, Cambridge, MA.