Corpus-Based Research on Chinese Language and Linguistics
edited by Bianca Basciano, Franco Gatti, Anna Morbiato

Form and Meaning Representation of Chinese Constructions
Fundamental Issues on Constructicography

Weidong Zhan (Peking University, China)

Jiajun Wang (Peking University, China)

Long Chen (Peking University, China)

Haibin Huang (Peking University, China)

Abstract This paper introduces a Chinese constructicon (CCL-CxnBank) and a corpus annotation platform for the description of actual usages of constructions in contexts. CCL-CxnBank is an online repository that contains more than 1,000 constructions, as well as the linguistic descriptions of their various features. Based on our practice of constructicography, we hold that constructions differ from phrases in that they are not recursive. We propose that the formal representation of a given construction should be linear, while its meaning should be represented through paraphrase templates and semantic frames. In the future, contextual features will be integrated to analyse the semantics of constructions.

Keywords Chinese constructicon. Constructicography. Construction grammar. Form and meaning representation. Principle of compositionality. Language engineering.

Edizioni Ca'Foscari

Sinica venetiana 6
e-ISSN 2610-9042 | ISSN 2610-9654
ISBN [ebook] 978-88-6969-406-6 | ISBN [print] 978-88-6969-407-3

Peer review | Open access **305**
Submitted 2020-10-02 | Accepted 2020-11-30 | Published 2020-12-21

DOI 10.30687/978-88-6969-406-6/010

1 Introduction

This paper introduces the work on knowledge representation of Chinese constructions done in recent years by the Centre for Chinese Linguistics (CCL) of Peking University. Our work includes two parts: the development of a Chinese constructicon (provisionally named as CCL-CxnBank)[1] and the annotation of a corpus consisting of sentences that display various usages of construction instances.[2] Our work stems from the belief that linguistic knowledge resources can better support natural language processing and language teaching if they are well organised, analysed, and digitised into databases and annotated corpora.

In the past 30 years, the construction approach to language has thrived among Chinese linguistic studies and has brought rich knowledge to both case studies and systematic studies (Zhang B. 2008, 2018; Zhang J. 2013). Against this background, since 2015 CCL has been running a project on the development of a Chinese constructicon database, which is the first Chinese constructicon project comprising both a construction knowledge database and an annotated corpus. CCL-CxnBank serves as a supplement to the current natural language engineering practice that in mainstream computational linguistics is based on commonly-used grammatical units, such as words and phrases. Up to now, this project has already collected over 1,000 Chinese constructions and recorded their syntactic, semantic, and pragmatic information. Moreover, relationships among constructions, such as synonymy, antonymy, and hyponymy/hyperonymy relations, have also been included, in order to provide a more systematic and coherent knowledge representation scheme for Chinese constructions. Finally, an online corpus annotation platform has been developed to annotate the internal structure and the subjective attitude meaning of each construct that occurs in real texts, with the aim of providing a comprehensive description of the actual usages of constructions in real contexts.[3]

This paper presents our work in progress and some of the major challenges we encountered in the development of CCL-CxnBank. § 2 presents our definition and understanding of the term 'construction' by comparing it with the conventional grammatical unit notion of 'phrase', which is commonly used to refer to a formal representation

1 The website of CCL-CxnBank is http://ccl.pku.edu.cn/ccgd.

2 We have also set up a website as a working platform for annotating the corpus, which is currently only accessible to authorised annotators. The website is http://162.105.161.162:8088/cclannotator/public/index.php.

3 'Construction' and 'construct' in this paper are used to refer to construction type and token respectively. 'Constructicon' refers to the construction database in which construction entries and their linguistic attributes are systematically organised and recorded.

scheme in syntactic structures in the knowledge engineering practices for computer. § 3 discusses issues in the representation of the forms and meanings of constructions. § 4 gives an overview of CCL-CxnBank and discusses the methodology adopted in its development. § 5 presents our work on corpus annotation, including an introduction of the online platform for annotation and some related challenges. The last section concludes by presenting the significance of our work and the future direction of development of construction resources.

2 The Properties of Constructions. Comparing Constructions with Phrases

From the viewpoint of language resources development, Zhan (2017) analysed the relationship and differences existing between constructions and conventional grammatical units, i.e. words, phrases etc. This work adopts Zhan's (2017) perspective: below, we discuss some major tenets and propose some further considerations.

Unlike some constructionists who maintain that all units of a grammatical system are constructions (Croft 2001), we treat constructions as complements to common phrases: in our view, constructions complement words and phrases rather than totally replacing them.[4] This is based on our understanding of constructions and conventional language units. Conventional language units can be classified into words and phrases. Words have fixed internal structures and cannot be recursively composed of smaller grammatical units. Phrases have expandable internal structures and can be recursively composed of smaller phrases. This classification allows greater efficiency and convenience in developing and maintaining language resource databases. In a language resource database, a limited (but large) number of words are listed entry by entry, while an infinite number of phrases can be described with a finite number of syntactic rules based on a finite number of grammatical categories such as noun, verb, noun phrase, verb phrase etc. However, in a linguistic system, other types of linguistic units can be identified (that we call 'constructions', Zhan 2017), which differ in the following respects.

First, constructions emerge from common phrases, which are formed by words. Therefore, constructions are different from words,

4 Treating words as constructions is merely a theoretical or labelling issue. Words *can* be treated as constructions from a 'form-meaning' pair perspective, but it makes little difference in the knowledge engineering practice. For languages with little or no inflection such as Chinese, knowledge in a dictionary is stored in exactly the same way as in constructions' description: each entry is a 'word form-word meaning' pair. In other words, referring to words as 'word constructions' or 'words' makes no difference in the knowledge engineering practice.

which are not composed of smaller grammar units. From the point of view of formal grammar, a word can even be regarded as the smallest grammatical unit or atomic unit and there is no need to analyse its internal components.

Second, constructions are different from phrases. In traditional linguistics, phrases are treated as core grammatical units. The formalisation of phrases includes four elements: relationships, heads, categories and hierarchies. These four elements jointly display a syntagmatic and recursive nature within phrases: (1) the syntagmatic relations between constituents within phrases, (2) the head roles in the phrases, (3) the grammatical categories the phrases and their constituents belong to, and (4) the hierarchical (tree) structures in which the phrases are internally organised. The syntactic description of these four aspects is the foundation for the computation of the meaning of phrases (Jurafsky, Martin 2000, chs. 15.1, 15.2). On the contrary, typical constructions have weak relationships between the constituents, no prominent head roles, only limited variations in their de-categorised components, and a linear internal structure rather than a hierarchical one. From the perspective of meaning, the acquisition of the meanings of phrases generally follows the so-called 'principle of compositionality', stating that the meaning of a whole sentence is acquired by the semantic combination of its constituent parts (Partee 2004). As for constructions, the meaning of a construct is the combination of the meanings of its constituents and the meaning of the construction in which these words occur. Therefore, constructions are not conventional phrases.

Third, we can either refer to constructions as phrases or refer to phrases as constructions (Croft 2001). If we refer to constructions as phrases, constructions are unique phrases; if we refer to phrases as constructions, phrases are schematised constructions (Zhan 2017). It is theoretically reasonable to refer to phrases as constructions; however, categorising them as the same grammatical unit does not mean they have identical grammatical properties. Constructions and conventional phrases still differ in many basic grammatical properties such as recursiveness and compositionality. For example, constructions can usually be embedded in conventional phrases, while only a limited number of phrases can be embedded into constructions. Example (1) illustrates two sentences with the same pattern: [不是 *búshì* + N_1 + 的 *de* + N_2].[5] N_1 differs from N_2 in (1a), while in (1b) N_1 and N_2

5 The glosses follow the general guidelines of the Leipzig Glossing Rules. Additional glosses include: BEI = 'Chinese 被 *bèi* marker', often labelled as a passive marker; DE = 'Chinese particle 的 *de*', functioning as modification marker or nominaliser; MP = 'mood particle' (in Chinese they are used to add various moods, including interrogation, request, command, emphasis and exclamation, to an utterance); SFP = 'sentence final particle'. In-text abbreviations are as follows: N = 'noun'; NP = 'noun phrase'; V = 'verb';

are identical (repetition of nouns with the same form): (1b) includes a construct of the construction [不是 *búshì* + N + 的 *de* + N], meaning 'N that is not N'.

1. a. 怎么解决这不是正品的问题?
zěnme jiějué zhè bú shì zhèng-pǐn de wèntí
how solve this not COP genuine-product DE problem
'How to solve the problem that this commodity is not genuine?'

   b. 怎么解决这不是问题的问题?
zěnme jiějué zhè bú shì wèntí de wèntí
how solve this not COP problem DE problem
'How to solve this problem which is not a problem?'

By comparing the examples above, it is obvious that the instance of the linear pattern [不是 + N1 + 的 de + N2] in (1a) has a different internal hierarchical structure, which can be expanded into a different form. (2) is the expansion of (1a), which maintains the original hierarchical structure.

2. 怎么解决这个商品不是厂家正品的严重失信问题?
zěnme jiějué zhè ge shāngpǐn bú shì chǎngjiā
how solve this CLF commodity not COP manufacturer
zhèng-pǐn de yánzhòng shīxìn wèntí
genuine-product DE serious dishonesty problem
'How to solve the problem that this commodity is not a genuine product of the manufacturer, which indicates a serious dishonest conduct?'

However, the instance of the pattern [不是 *bú shì* + N + 的 *de* + N] 'N that is not N' in (1b) cannot be expanded as that in (1a). [不是 *bú shì* + 问题 *wèntí* + 的 *de* + 问题 *wèntí*] 'a problem which is not a problem' is a fixed language unit: 问题 *wèntí* 'problem' can only be substituted with a limited number of nouns such as 办法 *bànfǎ* 'method', 理由 *lǐyóu* 'reason', 机会 *jīhuì* 'chance', 结局 *jiéjú* 'outcome', 妈妈 *māma* 'mother' etc. The generative capacity of this pattern is limited if compared with that of phrase patterns shown in example (1a) and (2). Furthermore, it carries an additional inherent meaning that goes beyond the meaning of 不是 *bú shì* and 问题 *wèntí*, which could be paraphrased as 'it is only a titular N' or 'it is not a typical N, but, nonetheless, we can grudgingly treat it as one' etc. The specific meaning is determined by the context in which the pattern occurs.

VP = 'verb phrase'; A = 'adjective'; AP = 'adjective phrase'; CLP = 'numeral plus classifier phrase'; X, Y,... = 'constituents with arbitrary syntactic category'.

Above all, constructions are different from conventional grammatical units, i.e. words and phrases, in a major respect: in language engineering, mapping between forms and meanings of constructions need to be listed entry by entry, just like those of words; combinatorial properties of constructions, on the other hand, need to be described like those of phrases.

3 The Form and Meaning Representation of Constructions

According to Zhan (2017) and following the considerations above, the forms of constructions should be described as linear patterns with specific lexical elements (which we call 'constants') and schematic elements (which we call 'variables'). Within a construction, constants are specific words, and variables are represented by part-of-speech tags or syntactic categories of phrases (N, V, NP, VP etc.). Some variables in certain constructions can be instantiated with elements of different phrase categories, which is mentioned above as 'de-categorisation'. The following examples illustrate the variables instantiated by word categories, phrase categories and cross-category elements.

Table 1 Some examples of constructions combined with constants and variables

Constructions	Constructs	Constants (Words)	Variables (Categories)
V + 一 *yī* 'one' + CLF + 是 *shì* 'COP' + 一 *yī* 'one' + CLF 'Every behaviour which V indicates counts'	说一句是一句 *shuō yī jù shì yī jù* 'Every word counts'	一 *yī* 'one' 是 *shì* 'COP'	V, CLF
说 *shuō* 'speak' + VP + 就 *jiù* 'immediately' + VP 'Carry out the behaviour indicated by VP immediately after promising to VP'	说给钱就给钱 *shuō gěi qián jiù gěi qián* 'Pay immediately after promising to pay'	说 *shuō* 'speak' 就 *jiù* 'immediately'	VP
除了*chúle* 'besides' + X + 还是 *háishì* 'still' + X 'There is nothing but X'	除了下雨还是下雨 *chúle xià yǔ háishì xià yǔ* 'It rains endlessly' 除了馒头还是馒头 *chúle mántou háishì mántou* 'There is nothing to eat but steamed buns'	除了 *chúle* 'besides' 还是 *háishì* 'still'	X (N, V, A etc.)

Constructions share semantic properties both with words and with phrases. On the one hand, the meanings of constructions have to be listed entry by entry just like words, in order to describe fixed relations between form and meaning. On the other hand, the meaning of constructions has to be computed by combining the meanings of the constituents following the 'principle of compositionality', just like phrases. The following two sections present and discuss issues in the representation of construction forms and meanings.

3.1 The Representation of Forms. Variations and Extensions of Constructs in Actual Use

The internal structure of a construction is represented as a linear pattern consisting of several constants and variables. While it is generally not necessary to consider recursiveness in the structural representation of a construction (typically, a construct cannot be embedded into a construct of the same construction), some constructions display a limited expansion capacity. Zhan (2017) analysed the basic forms of constructions, which are considered to be stable and fixed. Here we further discuss the form variations of constructions, which can be distinguished into three types.

3.1.1 The Variation of Lexically Specified Elements of a Construction

Let us consider the following examples:

3. a. 有什么大惊小怪的
yǒu shénme dàjīngxiǎoguài de
have what fuss DE
'There is nothing to fuss about'

b. 有什么可大惊小怪的
yǒu shénme kě dàjīngxiǎoguài de
have what may fuss DE
'There is nothing to fuss about'

c. 有什么好大惊小怪
yǒu shénme hǎo dàjīngxiǎoguài
have what worth fuss
'There is nothing to fuss about'

d. 有什么好大惊小怪的
yǒu shénme hǎo dàjīngxiǎoguài de
have what worth fuss DE
'There is nothing to fuss about'

a'. 没有什么大惊小怪的
méi yǒu shénme dàjīngxiǎoguài de
not have what fuss DE
'There is nothing to fuss about'

b'. 没有什么可大惊小怪的
méi yǒu shénme kě dàjīngxiǎoguài de
not have what may may fuss DE
'There is nothing to fuss about'
没有什么好大惊小怪

c'. *méi yǒu shénme hǎo dàjīngxiǎoguài*
not have what worth fuss
'There is nothing to fuss about'

d'. 没有什么好大惊小怪的
méi yǒu shénme hǎo dàjīngxiǎoguài de
not have what worth fuss DE
'There is nothing to fuss about'

The form of the construction in example (3) is [有 *yǒu* + 什么 *shénme* + VP + 的 *de*] 'there is no need to VP', as in (3a). (3b)-(3d) are variations of this construction with other constants added, such as 可 *kě* 'may', 好 *hǎo* 'worth', or with the constant 的 *de* omitted. 有 *yǒu* 'have' in these constructs may also appear in its negated form, 没有 *méi yǒu*, as in (3a')-(3d'), meaning 'there is no need to VP', 'it is worthless to VP' etc. The variations of a construction form can be either exhaustively listed in the constructicon or captured by regular expressions. The construction form in example (3) can be represented as [(有 *yǒu* | 没有*méi yǒu*) 什么 *shénme* (好*hǎo* | 可 *kě*)? (VP) (的*de*)?], where '?' indicates zero or one leftward character, and '|' indicates disjunction, matching either left or right character. Regular expressions can be represented by the finite state transition network (FSTN). The FSTN of the construction in example (3) is illustrated in figure 1 below (Chomsky 1956).

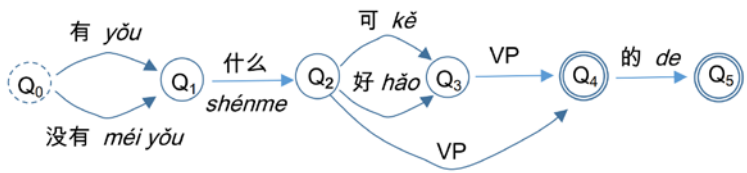

Figure 1 The FSTN recognising form variations of the construction [有 *yǒu* + 什么 *shénme* + VP + 的 *de*]

Among the 1,066 entries in CCL-CxnBank, 816 are marked as not having form variations, and 250 entries are marked as having some (about 23.45%). Constructions vary both in the number and the degree of form variations. The basic form of the construction in example (4) is [A + 就 *jiù* 'exactly' + A + 在 *zài* 'on' + X] 'it is indeed X which makes it A', with more complicated instantiations than example (3): (4a) is an instance that can match the construction form exactly; in (4b), the auxiliary 可能 *kěnéng* 'may' is inserted before 就 *jiù* 'indeed' as a constant of the construction; in (4c) and (4d), 就 *jiù* 'indeed' is replaced by 就是 *jiùshì* 'exactly' and 也就 *yě jiù* 'also exactly', respectively. Besides, in (4c) and (4d), the first variable is separated from the rest by a comma.

3. a. 个人申请贷款, 麻烦就麻烦在担保和抵押。
gèrén shēnqǐng dàikuǎn máfan jiù máfan
individual apply.for loan troublesome indeed troublesome
zài dānbǎo hé dǐyā
on guarantee and mortgage
'The troubles of individual application of loans lie exactly on guarantees and mortgages'.

b. 他的主子大人将来倒霉可能就倒霉在狗的身上。
tā de zhǔzi dàrén jiānglái dǎoméi kěnéng jiù
he DE master lord future unfortunate may indeed
dǎoméi zài gǒu de shēn shàng
unfortunate on dog DE body on
'Something unfortunate may happen to his lord master exactly because of the dog'.

c. 很多学生觉得文言文难, 就是难在一些实词和虚词上。
hěnduō xuéshēng juéde wényánwén nán jiùshì
many student think Classical.Chinese difficult exactly
nán zài yìxiē shící hé xūcí shàng
difficult on some content.word and function.word above
'Many students think that the difficulties of Classical Chinese lie exactly on some content words and function words'.

d. 处方的'含金量'高, 也就高在用进口药和合资企业药的比重猛增。
chùfāng de hánjīnliàng gāo yě jiù gāo
prescription DE gold.content high also indeed high
zài yòng jìnkǒu yào hé hézī qǐyè
on use imported medicine and joint.venture enterprise
yào de bǐzhòng měngzēng
medicine DE ratio soar
'The 'true value' (price) of the medical prescriptions is high exactly because of the soaring of the ratio of the medicines used, which are produced by foreign and joint venture enterprises'.

The form variations in (4a) and (4b) are complete grammatical units, while in (4c-d) the construction variations may not be grammatical constituents. In (4c), 难, 就是难在一些实词和虚词上 *nán jiùshì nán zài yìxiē shící hé xūcí shàng* 'difficulties lie on some content words and function words' can be treated either as a complete constituent or as two clauses separated by a comma, with each clause acting as a constituent. Thus, (4c) is no longer appropriate to be treated as a construct instantiated from the form variation of the construction [A + 就 *jiù* + A + 在 *zài* + X], at least not the same as that instantiated by examples (4a) and (4b), even though they almost share the same meaning.

This construction has even more form variations, such as (4a') and (4c') below, which are expanded from (4a) and (4c).

4. a'. 个人申请贷款的麻烦, 最主要就麻烦在担保和抵押。
gèrén shēnqǐng dàikuǎn de ***máfan*** *zuì zhǔyào*
individual apply.for loan DE trouble most main
jiù máfan zài dānbǎo hé dǐyā
indeed troublesome on guarantee and mortgage
'The troubles of individual application of loans lie exactly on guarantees and mortgages'.

c'. 文言文难, 很多学生觉得就是难在一些实词和虚词上。
wényánwén ***nán*** *hěnduō xuéshēng juéde* ***jiù shì***
Classical.Chinese difficult many student think indeed COP
nán zài yìxiē shící hé xūcí shàng
difficult on some content.word and function.word above
'The difficulties of Classical Chinese, some students think, lie exactly on some content words and function words'.

In (4a') and (4c'), if the bold parts are treated as the form variations of the construction [A + 就 *jiù* + A + 在 *zài* + X], some problems arise when trying to represent the form of the construction variations, because regular expressions will capture chunks with no linguistic significance when trying to match the constructs in the sentences. The chunks 最主要 *zuì zhǔyào* 'the most important' in (4a') and 很多学生觉得 *hěnduō xuéshēng juéde* 'many students think that...' in (4c') appear between a constant and a variable. A module needs to be specifically designed to handle these strings appropriately.

Examples in (3) and (4) show that, while the constants in [有 *yǒu* + 什么 *shénme* + VP + 的 *de*] have limited form variations which can be captured rather precisely and exhaustively by regular expressions, the relation between the first variable 'A' and the constant 就 *jiù* in [A + 就 *jiù* + A + 在 *zài* + X] is relatively loose. In real texts, language chunks of various categories can be inserted between the constant and the variable in the constructs, displaying great variability. Although these constructs express the same basic meaning, their forms cannot be exhaustively and appropriately described. The internal structure of the construction in (4) requires further examination. In other words, the construction [A + 就 *jiù* + A + 在 *zài* + X] is not a monolithic whole. The chunk responsible for the explanation is [A + 在 *zài* + X], occurring after 就 *jiù*. [就 *jiù* + A + 在 *zài* + X] is a relatively independent chunk, which can be used separately from the preceding variable 'A', as in (4a') and (4c'). It would be more reasonable to include [A + 在 *zài* + X] as a separate construction entry, specifying that it is a synonym of [A + 就 *jiù* + A + 在 *zài* + X]. When processing sentences with such constructs, the construct [A + 就 *jiù* + A + 在 *zài* + X] has the priority over the others, according to

the greedy matching principle. If [A + 就 *jiù* + A + 在 *zài* + X] fails to match any construct, [A + 在 *zài* + X] will be called in for matching.[6]

3.1.2 Expansion by Juxtaposition of Constructions in the Form of Chunks

Let us consider the following examples:

5. 新就新在财政部门认真对待全国人大、政协、‘两会’代表、委员的意见上，新在他们转变作风、行动迅速上。
xīn jiù xīn zài cáizhèng bùmén rènzhēn duìdài
new exactly new on financial department seriously treat
quánguóréndà zhèngxié liǎng-huì dàibiǎo
N.P.C. N.P.P.C.C. Two-Sessions representatives
wěiyuán de yìjiàn shàng xīn zài tāmen zhuǎnbiàn
committee DE advice on new be.at they change
zuòfēng xíngdòng xùnsù shàng
style act rapidly above
‘The novelty lies in the fact that the Financial Department took very seriously the advice given by N.P.C., N.P.P.C.C., and the representatives and committees of the Two Sessions, regarding their change in working style and action speed’.

6. 一整天在湖上晃呀晃、拐呀拐的也是一种度日方式吧。
yì zhěng tiān zài hú shàng huàng ya huàng
one whole day on lake on waggle MP waggle
ya guǎi de yě shì yì zhǒng dù rì fāngshì ba
MP turn DE also COP one CLF spend day way MP
‘Waggling and turning around all day long on the lake is also a way to spend the day’.

7. 喜剧不是喜剧, 闹剧不是闹剧, 丑角不是丑角, 痞子不是痞子, 简直滑稽至极。
xǐjù bú shì xǐjù nàojù bú shì nàojù
comedy not COP comedy farce not COP farce
chǒujué bú shì chǒujué pǐzi bú shì pǐzi
clown not COP clown ruffian not COP ruffian
jiǎnzhí huájī zhì jí
simply ridiculous to utmost

6 Another method is to treat examples (4a’) and (4c’) as separable usages of a construction, which requires form matching of a discontinuous string, thus making the matching process more complicated.

'It is neither a comedy nor a farce, the clown is not a clown and the ruffian is not a ruffian: it's ridiculous'.

The basic form of the construct in (5) is [A + 就 *jiù* + A + 在 *zài* + X] (same as in example (4)), with [A + 在 *zài* + X] partially expanding, appearing twice in the sentence. Similarly, the pattern [V + 呀 *ya* + V] 'V again and again', whose basic form is [V + 呀 *ya* + V + 的 *de*] 'V-*ing* again and again', expands and appears twice in (6). The basic form of the construction in (7) is [N_1 + 不是 *bú shì* N_1, N_2 + 不是 *bú shì* + N_2] 'it is neither N_1 nor N_2', which already includes two juxtaposed chunks. In (7), the whole construct expands, differently from (5) and (6), where the constructs only partially expand.

The 'Expandable' (是否可扩展 *shìfǒu kě kuòzhǎn*) feature in CCL-CxnBank is used to describe the constructs illustrated above. Its default value is 'true', which allows expansion by juxtaposition. For constructions which cannot expand juxtapositionally, the value will be 'false'.

8. a. 一个不留神, 摔了个大跟头。
yí ge bù liúshén shuāi le ge dà gēntou
one CLF no caution fall PFV CLF big somersault
'Without caution, (someone) fell heavily'.

b. 一个愿打, 一个愿挨。
yí ge yuàn dǎ yí ge yuàn āi
one CLF willing beat one CLF willing endure
'One is willing to beat, the other is willing to be beaten'.

c. 一个使劲骂一个偷东西的孩子, 还有一个 […]
yí ge shǐjìn mà yí ge tōu dōngxi de
one CLF continuously scold one CLF steal thing DE
háizi hái yǒu yí ge
child also have one CLF
'One keeps on scolding a child who steals, the other […]'

一个不留神 *yí ge bù liúshén* in (8a) is an instantiation of the construction [一 *yí* + 个 *ge* + VP] 'one moment of VP (leads to)...', which is also shared by instantiations such as [一 *yí* 'one' 个 *ge* 'CLF' 没 *méi* 'not' 站稳 *zhàn-wěn* 'stand-steady'] 'one moment of instability...', [一 *yí* 'one' 个 *ge* 'CLF' 手 *shǒu* 'hand' 软 *ruǎn* 'soft'] 'one moment of loosened grip...' etc., all conveying the happening of unexpected events which bring about undesirable results. However, although the sentences in (8b) and (8c) formally display the [一 *yí* + 个 *ge* + VP] pattern, they are not instantiations of this construction. Rather, 一个 *yí ge* 'one' acts as the subject (with the head noun omitted) of the following predicate. In (8c), there is also a second 一个 *yí ge*, which is

part of the modifier of the NP's head noun 孩子 *háizi* 'child', together with the relative clause 偷东西的 *tōu dōngxi de* 'who steals', altogether meaning 'the child who steals'. In CCL-CxnBank, the 'Expandable' feature of [一 *yí* + 个 *ge* + VP] is thus set to 'false', therefore preventing the chunks like those in (8b) and (8c) from being recognised as constructs of the construction [一 *yí* + 个 *ge* + VP] in automatic syntactic parsing.

3.1.3 Schematic Elements (Variables) Which May not Form a Constituent as a Whole

Let us consider the following examples:

9. a. 这批货要多少有多少。
zhè pī huò yào duōshǎo yǒu duōshǎo
this CLF goods require how.many have how.many
'As for this batch of goods, you can have as many as you need'.

b. 接下来不作解释了, 能理解多少理解多少。
jiēxiàlái bú zuò jiěshì le néng lǐjiě duōshǎo
next not conduct explain SF can comprehend how.much
lǐjiě duōshǎo
comprehend how.much
'(I) shall explain no more. Try to comprehend as much as (you) can'.

c. 观众爱给多少给多少, 不给也无妨。
guānzhòng ài gěi duōshǎo gěi duōshǎo
audience like give how.much give how.much
bù gěi yě wúfáng
not give also acceptable
'Audience may give as much as they like, even nothing'.

d. 有多少根发梢便会传递多少缕柔情蜜意。
yǒu duōshǎo gēn fà-shāo biàn huì chuándì
have how.many CLF hair-end therefore can convey
duōshǎo lǚ róuqíng-mìyì
how.many CLF tender-affection
'Men will be fascinated by her thick hair'.

(9a) displays the construction [V_1 + 多少 *duōshǎo* + V_2 + 多少 *duōshǎo*] 'the amount of V_1 leads to the same amount of V_2'. Chunks with similar patterns also appear in (9b)-(9d), which convey similar meanings, indicating that the quantity involved in the latter event is dependent on the quantity involved in the former one. However, chunks in (9b-d) cannot be treated as true instantiations of the construction

[V_1 + 多少 *duōshǎo* + V_2 + 多少 *duōshǎo*] 'the amount of V_1 leads to the same amount of V_2', in that the chunks after 多少 *duōshǎo* 'how many', such as [有 *yǒu* 'have'... 根 *gēn* 'CLF' 发梢 *fàshāo* 'hair end'] in (9d), do not form complete constituents. To account for this, sentences in (9) can be first treated with common phrase structure rules. Each sentence consists of two juxtaposed phrase structures and interrogative chunks with the same form generally occur at the same syntactic position. The whole structure expresses a dependency correlation, which can be instantiated by any number of event pairs with conditional relation. The quantity included in the second event corresponds to the quantity included in the first event.

Similar phenomena are more common in compound sentences. Take the construction [再 *zài* 'again' + VP_1 + 也 *yě* 'still' + VP_2] 'no matter how much one VP_1, VP_2 still occurs' as an example. Simple constructs can be decomposed into the constants 再 *zài* and 也 *yě*, and two predicative variables. However, for more complicated constructs, the pattern […再 *zài*…也 *yě*…] establishes a long-distance relation which connects two clauses, as happens in (10):

10. 你奉献得再多, 那些人也觉得不够
nǐ *fèngxiàn* *de* *zài* *duō* *nàxiē*
you give COMP again much those
rén *yě* *juéde* *bú* *gòu*
people still think not enough
'No matter how much you give, they will always think it is not enough'.

The meaning of the whole sentence can be decomposed into the basic propositional meanings of the two clauses with an adversative relation, which is represented by the two function words 再 *zài* and 也 *yě*. Describing the adversative relation using the linear pattern [再*zài* 'again' + VP_1 + 也 *yě* 'still' + VP_2] 'no matter how much one VP_1, VP_2 still occurs' is an over-simplification. In fact, the variables between 再 *zài* and 也 *yě* may not form a constituent, but separately belong to the two clauses as shown in example (10). In addition, the constant 也 *yě* can be replaced by other tokens, such as 都 *dōu*, 总 *zǒng*, 还 *hái* etc. (all roughly with the meaning 'still', when used here).

Constructions such as those in (9) and (10) require similar analyses: they are first processed using phrase structure rules and then marked as constructs with specific relations according to construction evoking elements such as [再 *zài*…也…*yě*], [多少 *duōshǎo*…多少 *duōshǎo*] etc.

3.2 The Representation of Meanings. A Strategy Combining Paraphrase Template and Semantic Frame

The semantics of common sentences follows the principle of compositionality: the meanings of words are combined according to the structural meanings of the sentences where these words occur, as is the case in (11).

11. 北大中文系培养计算语言学本科生
Běidà Zhōngwén-xì péiyǎng jìsuàn-yǔyánxué běnkē-shēng
PKU Chinese-department train computational-linguistics undergraduate-student
'The department of Chinese language and literature of PKU has an undergraduate program in computational linguistics'.

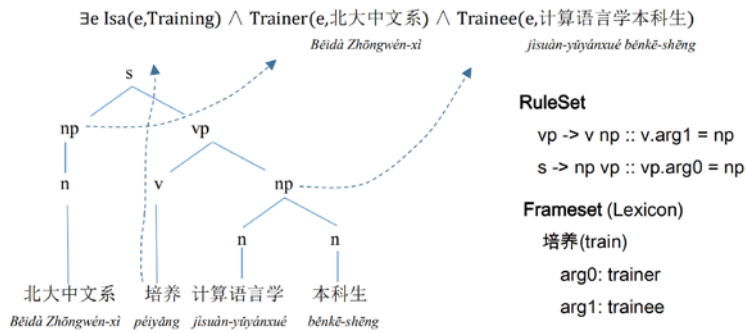

Figure 2 The semantic composition of sentence (11)

The syntactic structure derived from the syntactic rules set in (11) allows identification of the semantic roles of the NPs, where 北大中文系 *Běidà Zhōngwén-xì* 'the department of Chinese language and literature of PKU' plays the role of a 'trainer', which is often annotated as 'arg0' in propbank-style corpus, and 计算语言学本科生 *jìsuàn-yǔyánxué běnkē-shēng* 'undergraduates in computational linguistics' plays the role of a 'trainee', which is often annotated as 'arg1'.

One way to compute the meaning of a construct is to paraphrase it into a structure which can be handled by general phrase structure rules. The paraphrased sentence can then be processed by a semantic analyser, where a semantic representation can be computed according to the 'principle of compositionality'. See the example below:

12. 贝多芬十一岁时, 就已经显露了他的音乐天才, 被认为是莫扎特第二。
Bèiduōfēn shíyī suì shí jiù yǐjīng xiǎnlù le
Beethoven eleven year.old time then already show PFV
tā de yīnyuè tiāncái bèi rènwéi shì
he DE music talent BEI regard COP
Mòzhātè dì-èr
Mozart second
'Beethoven showed his music talent quite early, at the age of eleven. At that time, he was regarded as a second Mozart'.

莫扎特第二 *Mòzhātè dì-èr* 'a second Mozart' in example (12) is an instantiation of [N + 第二 *dì-èr*]. In the CCL-CxnBank, the 'Paraphrase Template'(释义模板 *shìyì múbǎn*) of this construction is set as either [像 *xiàng* + N + 一样 *yíyàng*] or [很 *hěn* + 像 *xiàng* + N], both meaning 'like N'. Thus, (12) can be paraphrased as 贝多芬十一岁时, 就已经显露了他的音乐天才, 被认为是很像莫扎特 *Bèiduōfēn shíyī-suì shí, jiù yǐjīng xiǎnlù-le tā de yīnyuè tiāncái, bèi rènwéi shì* ***hěn xiàng Mòzhātè*** 'Beethoven showed his music talent early at the age of eleven. At that time he was believed to be very much like Mozart', where 很像莫扎特 *hěn xiàng Mòzhātè* 'very much like Mozart' is an ordinary phrase structure, whose meaning can be computed by the semantic analyser designed for processing ordinary phrase structures.

The paraphrasing method encounters difficulties when dealing with complicated meanings of constructs, at least in the following two aspects. First, paraphrase templates fail in the constructs where there is a variable that does not form a constituent. The constructs illustrated in 3.1.3 with the pattern […再 *zài*…也 *yě*…], for example, display variables that do not form complete constituents. In this case, it is more appropriate to determine their meanings by first analysing the structure of the compound clauses where the construct appears, and then representing such meanings separately, rather than applying the paraphrasing method as in (12). Suppose there are two clauses S1 and S2, where S1 includes 再 *zài* and S2 includes 也 *yě*. The propositional meanings of S1 and S2 are separately represented as P1 and P2. The meaning of the whole sentence is represented with two predicate formulas 'AND(P1, P2)' and 'INEVITABLY(P2)', where the former represents the basic propositional meaning of the whole sentence, and the latter represents the subjective attitude brought by the constants 再 *zài* and 也 *yě*, expressing the speaker's attitude that P2 will inevitably happen.

Paraphrase templates also fail to process construction meanings when the acquisition of the meanings depend on the context rather than on the construction itself, such as [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N'. While the paraphrase templates of this construction is given in CCL-CxnBank, such as [凭借 *píngjiè* 'rely on' + N + 获得 *huòdé* 'gain' + 优势 *yōushì* 'advantage' / 认同 *rèntóng* 'approval' / 权力 *quánlì* 'power] 'gain advantage / approval / power

with N', the specific meaning of certain constructs has to be fully determined in the specific context.

More specifically, 说话 *shuō-huà* 'speak' may either have a literal meaning, as in [用智慧 *yòng zhìhuì* 'use wisdom' 说话 *shuō-huà* 'speak'], meaning 'speak with wisdom', or display a metaphoric reading, as in [用行动说话 *yòng xíngdòng* 'use action' + 说话 *shuō-huà* 'speak'], meaning 'speak with action', [用拳头说话 *yòng quántou* 'use fist' + 说话 *shuō-huà* 'speak'], meaning 'speak with fists'. The constant 说话 *shuō-huà* 'speak' in this construction can have different meanings in different contexts. Therefore, the meanings of the instances of this construction cannot be easily formalised through paraphrase templates, which can only provide abstract and general meaning descriptions. Some other representation schemas that try to represent construction meanings by paraphrasing also fail in this construction, e.g. AMR for constructions (Bonial et. al. 2018).

Construction meanings that are determined by context are more suitable to be formalised by frame representations, where constructional meanings can be included through attributes in the frame: specific meanings implied in certain contexts can be specified as values of the attributes. For example, the meaning of the construction [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N' can be represented with the frame in figure 3.

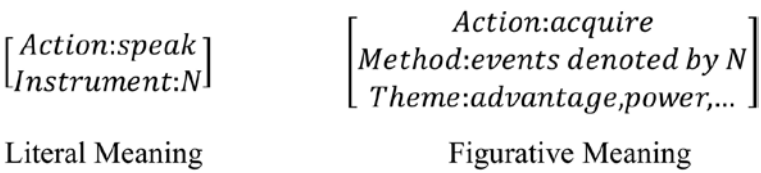

Figure 3 The frames representing the literal and figurative meanings of [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N'

The frames below represent the meaning of two instances of the construction: 用数据说话 *yòng shùjù shuō-huà* 'use figures speak, gain approval with data', and 用拳头说话 *yòng quántou shuō-huà* 'use fist speak, acquire power by beating others, assert one's authority through force'.

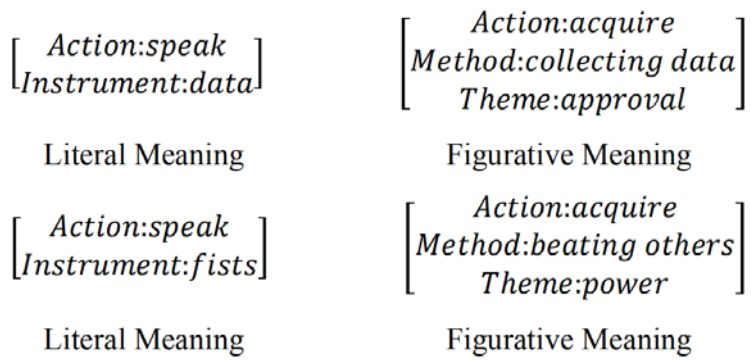

Figure 4 The frames representing the literal and figurative meanings of 用数据说话 *yòng shùjù shuō-huà* 'gain approval with data', and 用拳头说话 *yòng quántou shuō-huà* 'assert one's authority through force'

In conclusion, the meaning representation of constructions can be decomposed into two layers, including:

1. Paraphrase Template: the meanings which can be expressed by simply manipulating symbols linearly.
2. Semantic Frame: the implicit meaning of a construction often need to be represented with frames.

The semantic frame can be further divided into two types:

a. the Implied Meaning: additional semantic relations are acquired after the structural analysis of variables and constants, which are discontinuous and remotely related, expressing meanings such as the dependency meaning expressed by the repetition of interrogatives on the same syntactic position, or the adversative relation expressed by [再 *zài* + VP_1 + 也 *yě* + VP_2] 'no matter how much one VP_1, VP_2 still occurs' (see above), which further indicates a subjective attitude of necessity etc.

b. The Contextual Meaning: the meaning of a specific construct has to be clarified in the context where it occurs. For instance, in the construction [用 *yòng* 'use' + N + 说话 *shuō-huà* 'speak'] 'speak with N', the variable N expresses the means by which certain actions take place. The whole construction uses means as a metaphor of the purpose of an action, such as 用数据说话 *yòng shùjù shuō-huà* 'use data speak, support an idea' (lit. 'speak with data'), 用成绩说话 *yòng chéngjì shuō-huà* 'use grades speak, prove some ability' (lit. 'speak with grades'), and 用拳头说话 *yòng quántou shuō-huà* 'use fist speak, to defeat one's opponent' (lit. 'speak with fists') etc. The abstract

aspects of the construction's derived meanings (such as the abstract 'objective' meaning highlighted in certain constructions) can be described in the CCL-CxnBank, and the specific aspects, including subjective attitudes such as evaluations, standpoints and emotions etc., can be analysed and added according to the context while annotating the corpus. This aspect will be elaborated in § 5 below.

4 The Framework and Current Status of CCL-CxnBank

Some constructicon projects on several languages are described in Lyngfelt et al. (2018). The quantity of data included in these projects so far is not very large. A brief survey on these constructicons is listed in Appendix I. This section introduces the design framework and the current status of our project on the basis of Zhan (2017), where the basic issues of developing CCL-CxnBank were briefly introduced and discussed.

The descriptive framework of the construction knowledge is a core issue for the development of constructicons. Using the English rate-construction as an example, Fillmore, Lee-Goldman and Rhodes (2012) summarised six types of construction knowledge: (1) a bracketing formula with syntactic and semantic information attached to mother and daughter nodes; (2) a mnemonic name (used to address the constructions); (3) syntactic categories of the mother and daughter nodes, sometimes followed by informal descriptions of their syntagmatic distributions; (4) (optional) informal descriptions of the semantic information of the mother and daughter nodes; (5) an informal interpretation of the meaning of the construction as a whole (similar to traditional dictionary explanations); (6) annotated sentences containing the construction.

The German constructicography project described in Lyngfelt et al. (2018) concluded that, in order to appropriately describe the idiosyncratic characteristics of constructions of a specific language, the design of the description framework has to suit the grammatical characteristics of the specific target language, rather than trying to stipulate a universal grammatical framework for constructions of all the languages around the world. The design of the framework of CCL-CxnBank is in accordance with this view, implementing Yu (2003) and Zhan (1999; 2000) as the fundamental grammatical framework for the description of constructions, which have their origins in Zhu (1982; 1985).

Compared with Fillmore, Lee-Goldman and Rhodes (2012), we have developed a framework which allows to describe a richer amount of information in a more fine-grained manner (see Appendix II). The framework includes seven parts: (1) basic information,

(2) constants and variables, (3) relations between constants and variables, (4) syntactic information, (5) semantic information, (6) pragmatic information, (7) references. Each part describes a specific aspect of a construction entry. Due to space limitation, only the first part is explained and illustrated in detail below:[7]

- **Form Variations** (构式变体 *gòushì biàntǐ*) of a given construction carry the same variable(s) but different constant(s). For example, the construction [有 *yǒu* + 什么 *shénme* + VP + 的 *de*] 'what is the worth of VP' has form variations such as [有 *yǒu* + 什么 *shénme* + 可 *kě* + VP + 的 *de*], [有 *yǒu* + 什么 *shénme* + 好 *hǎo* + VP + 的 *de*] etc., as shown in § 3.1.1.
- **Construction Type** (构式类型 *gòushì lèixíng*) may either be fixed (凝固型 *nínggù xíng*), semi-fixed (半凝固型 *bàn-nínggù xíng*), phrasal (短语型 *duǎnyǔ xíng*) or compound (复句型 *fùjù xíng*). For example, [用 *yòng* 'use' + 脚 *jiǎo* 'foot' + 投票 *tóupiào* 'vote'] 'vote with feet' is a fixed construction, which contains only constants and no variable components. [人 *rén* 'person' + 见 *jiàn* 'meet' + 人 *rén* 'person' + V] 'whoever meets him/her V', on the other hand, is a semi-fixed construction: this type of construction usually has a fixed length, mostly four syllables, and contains one or two variants. [NP + 倒 *dào* 'but' 是 *shì* 'be' + NP] 'although NP is NP,...' is an example of phrasal constructions, which have variable length and contain more than one variable component; the variants are to some extent replaceable. Finally, [NP_1 + X_1, NP_2 + 还 *hái* 'still' + X_2 + 呢 *ne*] 'NP_1 X_1 is trivial compared with NP_2 X_2' is a compound construction, with variable length and more variants, which have higher replaceability. The definitions of these four types are elaborated in Zhan (2017).
- **Features** (构式特征 *gòu shì tèzhēng*): tags indicate construction features related to their syntax, semantics or other aspects. This set of tags is open, i.e. new tags can be added. For example, the tags attached to the construction [NP + 不 *bù* 'not' + VP + 谁 *shéi* 'who' + VP] 'NP do not VP, who else is supposed to do so' are: (i) 省略 *shěnglüè* 'ellipsis', as the original form of this construction is [如果 *rúguǒ* 'if' + NP + 不 *bù* 'not' + VP + 那么 *nàme* 'then' + 谁 *shéi* 'who' + VP], in which the conditional connectives 如果 *rúguǒ* 'if' and 那么 *nàme* 'then' marking the logical relation between the two clauses in the construction are omitted; (ii) 复现 *fùxiàn* 'recurrence', since there are two perfectly identical VPs in the construction; (iii) 含否定成分 *hán fǒudìng chéngfèn* 'containing negation markers', since there is a negative word 不 *bù* 'not' in the construction; (iv) 含疑问成分

7 For more details on the remaining six parts, please visit the website of CCL-CxnBank.

hán yíwèn chéngfèn 'containing question markers', since there is a question word 谁 *shéi* 'who'; and (v) 修辞 *xiūcí* 'rhetoric', since this construction is a rhetorical question.

- **Number of Syllables** (构式音节数 *gòu shì yīnjié shù*) captures the length of a construction or the number of syllables allowed for the construction. For fixed constructions it is a fixed number (e.g. (2) for [甩 *shuǎi* 'throw' + 锅 *guō* 'pot'] 'pass the buck'), while for other types of constructions it is a range of possible numbers (e.g. (4)-(10) for [有 *yǒu* 'have' + 什么 *shénme* 'what' + NP] 'there is no NP').
- **Number of Chunks** (组块数 *zǔ kuài shù*): the number of chunks of the construction that describes whether the construction is separated into two parts by a comma. For example, [没有*méi yǒu* 'not have' + NP_1 + 就 *jiù* 'then' + 等于 *děngyú* 'be equal to' + 没有 *méi yǒu* 'not have' + NP_2] 'the loss of NP_1 leads to the loss of NP_2' and [别 *bié* 'don't' + 说 *shuō* 'speak' + X, 连 *lián* 'even' + Y + 都 *dōu* 'all' + Z] 'even Y Z, let alone X' are both compound constructions, but the former has one chunk, while the latter has two chunks.
- **Expandable** (是否可扩展 *shìfǒu kě kuòzhǎn*) refers to the property of whether the construction can be expanded by juxtaposition. For example, [A + 得 *de* 'COMP' + 够呛 *gòuqiāng* 'terribly/extremely'] 'extremely A' can be expanded by juxtaposition, in sentences such as 他累得够呛，困得够呛，倒头就睡 *tā lèi de gòuqiàng, kùn de gòuqiàng, dǎotóu jiù shuì* 'he fell asleep immediately, as he was extremely tired and sleepy'.
- **Sense Number** (义项编号 *yìxiàng biānhào*) indicates the number of the meanings of a construction; not all constructions with the same form share the same meaning. If a construction form has only one meaning, the sense number of this entry is recorded as 0. Constructions with the same form but different meanings are listed in CCL-CxnBank as different entries, with sense numbers recorded as 1, 2, 3,... etc.
- **Paraphrase Templates** (释义模板 *shìyì múbǎn*) specifies the ordinary phrase that is synonymous to a construction. This column records phrases which can replace constructs of the entry in language use. For example, the paraphrase templates of [NP + 不 *bù* 'not' + VP + 谁 *shéi* 'who' + VP] 'NP do not VP, who else is supposed to VP' are: [如果 *rúguǒ* 'if' + NP + 不 *bù* 'not' + VP + 那么 *nàme* 'then' + 谁 *shéi* 'who' + VP] 'if NP does not VP, then who else is supposed to do so', [NP + 一定 *yídìng* 'surely' + 会 *huì* 'be likely' + VP] 'NP is surely to VP', [NP + 就 *jiù* 'then' + 该 *gāi* 'should' + VP] 'NP is obliged to VP' etc.
- **Samples** (构式实例 *gòushì shílì*) specify at least 3 samples of the actual usages of the construction, either in contexts or not. The

samples are collected from the CCL corpus[8] or built by the lexicographer according to her/his intuition. For example, the samples of [NP + 不 *bú* 'not' + VP + 谁 *shéi* 'who' + VP] 'NP does not VP, who else is supposed to VP' are: 劳模不干谁干 *láomó bù gān shéi gàn* 'if the model worker doesn't do it, who else is supposed to do it', 你不失败谁失败 *nǐ bù shībài shéi shībài* 'if you do not fail, who else is supposed to fail', and 我不入地狱谁入地狱 *wǒ bú rù dìyù shéi rù dìyù* 'if I do not step into hell, who else is supposed to do so'.

- **Synonym Constructions** (同义构式 *tóngyì gòushì*):[9] construction entries which share the same meaning and the same constants of the construction. For example, the synonym construction of [N_1 + 多 *duō* 'much, many' + N_2 + 少 *shǎo* 'few, little'] 'N_1 is abundant while N_2 is lacking' is [V_1 + 多 *duō* 'much' + V_2 + 少 *shǎo* 'little'] 'always V_1 but seldom V_2', and vice versa. The two constructions share the same template of interpretation.
- **Antonym Constructions** (反义构式 *fǎnyì gòushì*): construction entries which have the opposite meaning of the construction. For example, the antonym construction of [NP + 倒 *dào* 'but' 是 *shì* 'be' + NP] 'although NP is NP,...' is [NP + 倒 *dào* 'but' 不是 *bú shì* 'not be' + NP] 'although NP is not NP', and vice versa.
- **Hyperonym Constructions** (上位构式 *shàngwèi gòushì*) specifies the more general construction entry which subsumes (both syntactically and semantically) the construction.[10]
- **Hyponym Constructions** (下位构式 *xiàwèi gòushì*): the more specific construction entries which are subsumed (both syntactically and semantically) by the construction.
- **Negated Forms** (否定形式 *fǒudìng xíngshì*): the constructions collected in the CCL-CxnBank are idiosyncratic patterns which cannot be further decomposed with phrase structure rules. Therefore, their negated forms have to be manually recorded rather than deduced with phrase structure rules. The same goes for **Interrogative Forms** (疑问形式 *yíwèn xíngshì*). See examples (13) and (14) below for a comparison between a common sentence that has a corresponding interrogative form and a construction that has no corresponding interrogative form. For constructions which do not have negated forms or inter-

8 http://ccl.pku.edu.cn:8080/ccl_corpus.

9 Ideally, the information content of this field, including synonym, antonym, hypernym and hyponym, can help establish hierarchical network relationships between constructs. But, in fact, there are only some local relationships of parts of constructs at present, and no network relationships covering all the constructions has been established.

10 There is nothing to fill in the field 'Hyperonym Constructions' in the current database, since there is no schematic construction recorded in CCL-CxnBank at the current stage. The same goes for 'Hyponym Constructions'.

rogative forms, or which are already negated or interrogative, these two columns are recorded as 'none'.

- **Origin** (形成机制 *xíngchéng jīzhì*): describes how a construction emerges, or the process of grammatical constructionalisation of a construction. An academic paper is usually required to explain the origin of a construction.
- **Notes** (备注 *bèizhù*): the place where the lexicographer may keep notes on issues related to an entry, which need to be logged in detail for further investigation.

The goal of CCL-CxnBank is to accurately describe all the syntactic distribution information of each construction, which is illustrated in the following examples.

13. a. 张三也买了那本书。
Zhāngsān yě mǎi le nà běn shū
Zhangsan also buy PFV that CLF book
'Zhangsan also bought that book'.

b. 谁也买了那本书?
shéi yě mǎi le nà běn shū
who also buy PFV that CLF book
'Who also bought that book?'

c. 张三也买了哪本书?
Zhāngsān yě mǎi le nǎ běn shū
Zhangsan also buy PFV which CLF book
'Which book did Zhangsan also buy?'

14. a. 连张三也买了那本书。
lián Zhāngsān yě mǎi le nà běn shū
even Zhangsan also buy PFV that CLF book
'Even Zhangsan bought that book'.

b. *连谁也买了那本书?
lián shéi yě mǎi le nà běn shū
even who also buy PFV that CLF book
*'Even who also bought that book?'

c. *连张三也买了哪本书?
lián Zhāngsān yě mǎi le nǎ běn shū
even Zhangsan also buy PFV which CLF book
*'Even which book did Zhangsan also buy?'

(13a) is a sentence whose internal structure is subject-predicate, while (13b) and (13c) are its interrogative forms. In general, sentences consisting of regular phrases have both a declarative form and a corresponding interrogative form. However, (14a) is an instance of the [连 *lián* 'even' + X + 也 *yě* 'also' + Y] construction, which does not have interrogative forms like those of (13a). Both (14b) and (14c), which contain the question words 谁 *shéi* 'who' and 哪 *nǎ* 'which', respectively, are ungrammatical.

Based on the detailed description of each construction, a variety of statistical information on all entries in CCL-CxnBank is available now. There is a web page that displays the frequency of occurring constants, variables, and features, including both single features and combinations of features, which can be extracted from all the constructions or just only from a selected type of constructions. For example, figure 5 shows the 8 most frequently occurring constants in CCL-CxnBank. They are 不 *bù* 'not', 一 *yī* 'one, a', 的 *de* 'DE', '是 *shì* 'be', 个 *ge* 'CLF', 有 *yǒu* 'have', 了 *le* 'PFV', 也 *yě* 'also', in descending order of frequency. Obviously, high frequency function words and verbs with more abstract meanings are more common in constructions.

Figure 5 The webpage that displays statistics of CCL-CxnBank

The left side of figure 5 shows the statistical results, i.e. the frequency list of items being counted. The right side of figure 5 shows a menu for the user to select 'Items that need to be counted', 'Scope of statistics', which have been explained above, and 'Sort criteria' (the statistical result can be presented both in order of frequency or in alphabetical order).

Based on the statistics of variable components and features in current CCL-CxnBank, we can sketch an overview of common features of Chinese constructions: (1) the top three variable categories (ignoring the category X which matches all the categories) are V (verb), A (adjective) and AP (adjective phrase), indicating that predicative con-

stituents are more likely to fill the slots of constructions than nominal constituents; (2) the top three construction features are recurrence (复现 *fùxiàn*), grammatical mismatch (语法错配 *yǔfǎ cuòpèi*) and ellipsis (省略 *shěnglüè*), which conforms to our expectation that, according to phrase-based rules of grammar, Chinese constructions usually have grammatical mismatches to some extent, which are often caused by recurrence or ellipsis of certain constituents.

5 Building a Syntactically and Semantically Annotated Corpus of Chinese Constructions

5.1 An Online Platform for the Annotation of Constructs

As a hand-built knowledge base, CCL-CxnBank alone cannot fully reflect the constructs' overall usages in real texts, especially their form and meaning variations. Just as lexicons and phrase structure rule bases have to be accompanied by treebanks to reflect the overall usages of linguistic units, constructicons too have to be accompanied by annotated corpora, in which each construction entry is complemented with a collection of sentences where the corresponding constructs occur.

The English FrameNet constructicon described in Lyngfelt et al. (2018) contains 73 constructions and 1,471 annotated sentences. The constructs in the sentences are annotated with linguistic information, including construction elements (CE), construction-evoking elements (CEE), words in the sentence and their syntactic categories etc. The linguistic information annotated on the constructs are mainly concerned with the constituents of the constructs, and the direct analysis of the meaning of the constructs is lacking.

In order to fill this gap, i.e. to fully reflect the uses of constructions in real texts and to investigate the sentiment information carried by constructions (Huang, Zhan 2018), we have selected from CCL-CxnBank 50 constructions that have subjective attitudinal meanings. These constructions are tagged with construction features such as negative evaluation (负面评价 *fùmiàn píngjià*), subjective large amount (主观大量 *zhǔguān dàliàng*), and subjective little amount (主观小量 *zhǔguān shǎoliàng*) in the database table that describes the basic information of the construction. For each of the 50 constructions, about 100 sentences from the CCL corpus are extracted, resulting in a total of 4,777 sentences.

For constructs within sentences, three types of information are annotated: the construct's boundary, constituents, and the subjective attitudinal meaning. A construct's boundary serves to separate a construct from its surrounding context. Within the boundary, con-

stituents are respectively annotated as constants and variables, according to the pattern of the construction. In figure 7, the coloured tiles highlight the construct 别说干事业，连吃饭走道都打不起精神 *bié shuō gàn shìyè, lián chīfàn zǒudào dōu dǎ bù qǐ jīngshén* 'be spiritless even when walking and eating, let alone working' in its context 一个人要是没有奋斗目标 *yí ge rén yàoshi méiyǒu fèndòu mùbiāo* 'if a person does not have a goal to strive for', with black tiles indicating the constants and red tiles indicating the variables.

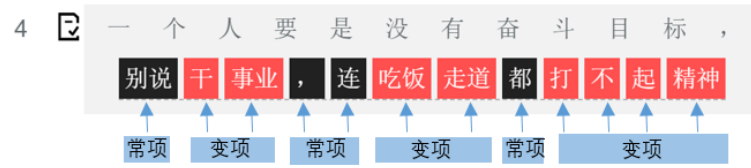

Figure 6 The annotation of the constituents of a construct. The constituents in black tiles and red tiles are constants and variables, respectively. The check mark on the top left corner indicates that this sentence's annotation has been proof-read.

As for the subjective attitudinal meaning, four dimensions are designed to describe it: evaluation (评价 *píngjià*), standpoint (立场 *lìchǎng*), emotion (情感 *qínggǎn*), and intensity (强度 *qiángdù*). As for evaluation, there are three options: positive (正面 *zhèngmiàn*), negative (负面 *fùmiàn*), or neutral (中立 *zhōnglì*). Standpoint also has three options to choose from: accept (接受 *jiēshòu*), refuse (拒绝 *jùjué*), or noncommittal (不置可否 *bù zhìxìn kěfǒu*). The value of emotion can be defined by the annotator according to her/his judgement on the emotion the specific construct expresses in the context. As for intensity, four values are given to choose from: none (缺省 *juéduì*),[11] very high (极 *jí*), high (很 *hěn*), or not high (不很 *bù hěn*). Below is the subjective attitudinal meaning of the construct 别说干事业，连吃饭走道都打不起精神 *bié shuō gàn shìyè, lián chīfàn zǒudào dōu dǎ bù qǐ jīngshén* 'be spiritless even when walking and eating, let alone working'.

Statistics of subjective attitudinal meanings are shown in table 2 below. Among the 4,777 sentences of 50 constructions, about 70% of them are concerned with evaluations and standpoints; about 25% of the sentences express emotions; about half of the sentences have a relatively high intensity of subjective attitudes.

11 'None' is the default option. It is used to check automatically whether the intensity of a sentence is marked or not by the platform.

Figure 7 The subjective attitudinal meaning of the construct 别说干事业, 连吃饭走道都打不起精神 *bié shuō gàn shìyè, lián chīfàn zǒudào dōu dǎ bù qǐ jīngshén* 'be spiritless even when walking and eating, let alone working', with evaluation set as 'negative', standpoint set as 'refuse', emotion set as 'detest', and intensity set as 'not high'

Table 2 Statistics of subjective attitudinal meanings in the annotated construction corpus

Evaluation		Standpoint		Emotion	Intensity			Total
positive	negative	accept	refuse		very high	high	not high	
1,141	2,223	1,204	2,111	1,238	785	1,731	1,022	4,777
23.89%	46.53%	25.20%	44.19%	25.92%	16.43%	36.24%	21.39%	
3,364		3,315			3,538			
70.42%		69.40%			74.06%			

The subjective attitudinal meaning, as its name implies, is subjective, and it is up to the annotator's language intuition to determine the value of the four dimensions, given a specific construct and its context. In this project, each construct is annotated by one annotator and checked by another annotator to ensure the internal consistency of annotation results, in order to control the quality of the annotation.

5.2 Some Challenges in the Annotation of the Form and Meaning of Constructs

The annotation of constructs is a challenging task in language resource development. There are several issues in annotating the forms and meanings of constructs. This is shown in the following example of the annotation of the [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] 'even X do/be Y' construction.

15. 连他离京, 做妹妹的都不知道。
lián tā lí Jīng zuò mèimei de dōu bù zhīdào
even he leave Beijing do sister DE all not know
'Even his sister does not know his departure from Beijing'.

In (15), the text string between the constants 连 *lián* and 都 *dōu* does not form a constituent, but stretches across two clauses. Therefore, the form [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] does not precisely match the construct in (15), which requires a more flexible representation of the form of the construction [连 *lián* 'even' + X + 都 *dōu* 'all' + Y]. It is the same situation as the one we have shown in example (10) for the pattern […再 *zài* 'again' …也 *yě* 'still'…]: the pattern [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] too establishes a long-distance relation which connects two clauses. In (15), the two clauses are 他离京 *tā lí Jīng* 'he leaves Beijing' and 做妹妹的不知道 *zuò mèimei de dōu bù zhīdào* 'his sister does not know', respectively, and are separated by a comma. The internal components of sentence (15) are analysed in the same way as sentence (10) in § 3.1.3.

16. 别说放弃了棋类的爱好, 连一般人天天都看的电视都没空看。
bié-shuō fàngqì le qílèi de àihào lián yìbān
not-say give.up PFV chess DE hobby even ordinary
rén tiāntiān dōu kàn de diànshì dōu méi-kòng kàn
person every.day all watch DE TV all not-time watch
'(He) does not even have time for TV programs that ordinary people watch, let alone having time for hobbies like playing chess'.

In (16), the second 都 *dōu* 'all' is a constant of the construction, but the first 都 *dōu* 'all' is used as a common adverb. This gives rise to difficulties when we try to design algorithms to automatically identify the construct's boundary.

17. 下雨天, 别说打(不)到车, 连地铁都会挤爆。
xiàyǔ tiān bié-shuō dǎ (bú) dào chē lián dìtiě dōu
rain day not-say call not able taxi even subway also
huì jǐbào
will overcrowded
'On rainy days, the subways will be crowded, not to mention that you cannot find a taxi'.

In (17), the speaker means that the hearer cannot find a taxi, and public transportation is not a solution, no matter whether the negative 不 *bù* appears in the clause introduced by 别说 *bié shuō* 'not to say' or not. This meaning is inferred from the literal meaning of the [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] construction. The mechanism of how a construct interacts with constituents outside of its boundary is a challenging problem and is still under investigation.

18. 这是连天气预报都可以放假的日子。
zhè shì lián tiānqì-yùbào dōu kěyǐ fàngjià de rìzi
this COP even weather-forecast all can have.a.day.off DE day
'The weather is so good that even the weather forecast can have a day off'.

In (18), the literal meaning 'the weather forecast being able to have a day off' is an improbable event. The occurrence of this improbable event is caused by the fact that the weather is extremely pleasant, so there is no reason to worry about changes in weather. The construction [连 *lián* 'even' + X + 都 *dōu* 'all' + Y] invites listeners to discover the reason for the occurrence of an improbable event. The mechanism by which this inference is carried out also needs further investigation.

The current construct annotation project is still in the early stages of exploration. Our goal is to annotate construction information based on treebanks and propbanks, where basic syntactic and semantic information has already been annotated. In this way, further investigation on the interaction between the constructs and the contexts can be carried out, where pragmatic information (such as inferences) shall be elicited and added into CCL-CxnBank.

6 Conclusions

As Ronald Langacker said in his book, "language is a mixture of regularity and idiosyncrasy" (1987, 411). During the development of Peking University Treebank (Zhan 2016), we already realised that constructions are necessary complements to common phrase structures, and common phrases are well suited to describe their internal constructs in terms of recursive tree structures defined by a formal grammar. However, for the constructions discussed in this paper, it is not suitable to describe their internal structures with hierarchical tree structures. As already pointed out in the analysis above, it is more suitable to describe the internal composition patterns of constructions as flat linear sequences.

The practical work of developing CCL-CxnBank taught us that constructicons and annotated construction corpora should be compatible with existing language resources, make full use of the work under the theory of phrase structure grammar, and integrate their annotation guidelines into systems of language resources such as treebanks, propbanks and FrameNet etc. The new language resources developed in this way will be more valuable from the perspective of language engineering.

As to the meaning representation of constructions, we recognise that, although constructional approaches to language emphasise the

integrity of constructions and neglect the combinatorial semantic analysis of the constituents of constructions to some extent, the principle of compositionality holds in the analysis of construction meanings. In order to correlate the form and meaning of a construction, it is still necessary to decompose the construction form and combine the meanings of the constituents. This principle deserves much consideration in the design of the annotation of construction constituents and meanings. On the other hand, another principle of semantic analysis, i.e. the contextuality principle, should also be considered in the analysis of construction meanings in our future research. The analysis of construction meanings needs to be combined with the annotation of contextual features of constructions.

Bibliography

Bonial, C. et al. (2018). "Abstract Meaning Representation of Constructions. The More We Include, the Better the Representation". *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC, May 2018, Miyazaki, Japan). Luxembourg; Paris: European Language Resource Association, 1677-84. http://www.lrec-conf.org/proceedings/lrec2018/pdf/856.pdf.

Chomsky, N. (1956). "Three Models for the Description of Language". *Transactions on Information Theory*, 2(3), 113-24. https://doi.org/10.1109/tit.1956.1056813.

Croft, W.; Cruse, D.A. (2004). *Cognitive linguistics*. Cambridge: Cambridge University Press.

Croft, W. (2001). *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Fillmore, C.J.; Kay, P.; O'Connor, M.C. (1988). "Regularity and Idiomaticity in Grammatical Constructions. The Case of Let Alone". *Language*, 64(3), 501-38. https://doi.org/10.2307/414531.

Fillmore, C.J.; Lee-Goldman, R.R.; Rhodes, R. (2012). "The FrameNet Constructicon". Boas, H.C.; Sag, I.A. (eds), *Sign-Based Construction Grammar*. Stanford, CA: CSLI Publications, 309-72.

Goldberg, A.E. (1995). *A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.

Goldberg, A.E. (2013). "Constructionist Approaches". Hoffmann, T.: Trousdale, G. (eds), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 15-31.

Hoffmann, T. (2017). "The Renaissance of Constructions. From Constructions to Construction Grammars". Dancygier, B. (ed.), *The Cambridge Handbook of Cognitive Linguistics*. Cambridge: Cambridge University Press, 284-309.

Huang S. 黄思思; Zhan W. 詹卫东 (2018). "Mianxiang qinggan fenxi de goushi zhuguan taidu yi chutan" 面向情感分析的构式主观态度义初探 (A Rudimentary Investigation of the Subjective Attitudinal Meaning of Construction Towards Sentiment Analysis). *Waiyu Jiaoxue*, 39(6), 27-33.

Jurafsky, D.; Martin, J.H. (2000). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Pearson Education.

Kay, P.; Fillmore, C.J. (1999). "Construction Grammar and Linguistic Generalizations. The What's X Doing Y? Construction". *Language*, 75(1), 1-33. https://doi.org/10.1353/lan.1999.0033.

Kay, P.; Michaelis, L.A. (2012). "Constructional Meaning and Compositionality". Maienborn, C.; von Heusinger, K.; Portner, P. (eds), *Semantics. An International Handbook of Natural Language Meaning*, vol. 3. Berlin: Mouton de Gruyter, 2271-96.

Langacker, R.W. (1987). *Foundations of Cognitive Grammar. Theoretical Prerequisites*, vol. 1. Stanford: California: Stanford University Press.

Li C.N.; Thompson, S.A. (1981). *Mandarin Chinese A Functional Reference Grammar*. Berkeley; Los Angeles: University of California Press.

Lyngfelt, B. et al. (eds) (2018). *Constructicography. Constructicon Development Across Languages*. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.22.

Partee, B.H. (2004). "Compositionality". Partee, B.H. (ed.), *Compositionality in Formal Semantics. Selected Papers*. Malden (MA): Blackwell, 153-81.

Yu S. 俞士汶 (2003). *Xiandai Hanyu yufa xinxi cidian xiangjie* 现代汉语语法信息词典详解 (The Grammatical Knowledge Base of Contemporary Chinese. A Detailed Explanation). Beijing: Tsinghua University Press.

Zhan W. 詹卫东 (1999). "Yi ge Hanyu yuyi zhishi biaoda kuangjia. Guangyi peijia moshi" 一个汉语语义知识表达框架: 广义配价模式 (A Framework of Chinese Semantic Representation. Generalised Valence Mode). *Proceedings of Joint Symposium on Computational Linguistics* (Beijing, 1-3 November 1999). Beijing: Tsinghua University Press, 1-7.

Zhan W. 詹卫东 (2000). *Mianxiang Zhongwen xinxi chuli de xiandai Hanyu duanyu jiegou guize yanjiu* 面向中文信息处理的现代汉语短语结构规则研究 (A Study of Constructing Rules of Phrases in Contemporary Chinese for Information Processing). Beijing: Tsinghua University Press.

Zhan W. 詹卫东 (2017). "Cong duanyu dao goushi. Goushi zhishiku jianshe de ruogan lilun wenti tanxi" 从短语到构式: 构式知识库建设的若干理论问题探析 (On Theoretical Issues in Building a Knowledge Database of Chinese Constructions). *Zhongwen Xinxi Xuebao*, 1, 230-8.

Zhang B. 张伯江 (2008). "Jushi yufa lilun yu Hanyu jushi yanjiu" 句式语法理论与汉语句式研究 (Constructional Approaches to Grammar and Research on Chinese Constructions). Sheng Y. 沈阳; Feng S. 冯胜利 (eds), *Dangdai yuyanxue lilun he Hanyu yanjiu* 当代语言学理论和汉语研究 (Contemporary Linguistic Theories and Chinese Linguistic Studies). Beijing: The Commercial Press, 497-507.

Zhang B. 张伯江 (2018). "Goushi yufa yingyong yu Hanyu yanjiu de ruogan sikao" 构式语法应用于汉语研究的若干思考 (Some Reflections on the Application of Construction Grammar in Chinese Studies). *Yuyan jiaoxue yu yanjiu*, 192(4), 2-11.

Zhang J. 张娟 (2013). "Guonei Hanyu goushi yufa yanjiu shi nian" 国内汉语构式语法研究十年 (Ten Years of Construction Grammar Research in China). *Hanyu Xuexi*, 2, 65-77.

Zhu D. 朱德熙 (1982). *Yufa jiangyi* 语法讲义 (Lecture Notes on Grammar). Beijing: The Commercial Press.

Zhu D. 朱德熙 (1985). *Yufa dawen* 语法答问 (Questions and Answers on Grammar). Beijing: The Commercial Press.

Appendix I. Constructicon Development across Languages

The table below is summarised from the content of each chapter in Lyngfelt et al. 2018.

Language	Name	Statistics	Website	Resources Dependent on
English	FrameNet Constructicon	73 constructions	http://www1.icsi.berkeley.edu/~hsato/cxn00/21colorTag/index.html	FrameNet Lexicon
Swedish	SweCcn	400 constructions	https://spraakbanken.gu.se/konstruktikon	Språkbanken SweFN++ Karp/Korp
Brazilian Portuguese	FN-Br	289 constructions	https://www.ufjf.br/framenetbr-eng/projects/frames-and-constructions http://webtool.framenetbr.ufjf.br/index.php/webtool/report/cxn/main	FrameNet
Japanese	JFN	N/A	http://jfn.st.hc.keio.ac.jp	FrameNet
Russian	FrameBank	Including 2700 high frequency verbs and 600 constructions which contain them	http://framebank.ru https://github.com/olesar/framebank	Språkbanken
German	GCon	39 constructions	http://gsw.phil.uni-duesseldorf.de https://gsw.phil.hhu.de https://gsw.phil.hhu.de/constructicon/constructionindex	FrameNet

Appendix II: The Framework of CCL-CxnBank