sciendo

Lingua Sinica
Vol.6, 2020
10.2478/linguasinica-2020-0004

Applying Collocation Analysis to Chinese Discourse: A Case Study of Causal Connectives

Yipu Wei[1], Dirk Speelman[2], Jacqueline Evers-Vermeul[3]
weiyipu@pku.edu.cn
[1]*School of Chinese as a Second Language, Peking University*
[2]*Research Unit of Quantitative Lexicology and Variational Linguistics, University of Leuven*
[3]*Utrecht Institute of Linguistics OTS, Utrecht University*

Abstract: Collocation analysis can be used to extract meaningful linguistic information from large-scale corpus data. This paper reviews the methodological issues one may encounter when performing collocation analysis for discourse studies on Chinese. We propose four crucial aspects to consider in such analyses: (i) the definition of collocates according to various parameters; (ii) the choice of analysis and association measures; (iii) the definition of the search span; and (iv) the selection of corpora for analysis. To illustrate how these aspects can be addressed when applying a Chinese collocation analysis, we conducted a case study of two Chinese causal connectives: *yushi* 'that is why' and *yin'er* 'as a result'. The distinctive collocation analysis shows how these two connectives differ in volitionality, an important dimension of discourse relations. The study also demonstrates that collocation analysis, as an explorative approach based on large-scale data, can provide valuable converging evidence for corpus-based studies that have been conducted with laborious manual analysis on limited datasets.

Keywords: Collocation Analysis, Word Associations, Discourse, Chinese Connectives

1 Introduction

An important advantage of linguistic corpora is that they allow linguists to examine naturally occurring data that are representative of the language population under investigation (McEnery and Hardie 2012). Modern corpora represent increasingly more genres and modalities and have seen an enormous increase in size. Techniques of corpus linguistics have also developed in recent decades, with increased input from computer science and statistics. These improvements in both corpora and methods have opened up new opportunities for discourse studies, which have largely relied on manual annotations and analysis.

Collocation analysis is a quantitative method for large-scale data analysis in corpus studies (Church et al. 1991; Church and Hanks 1990; Evert 2005, 2008; Manning and Schütze 2000; Stefanowitsch and Gries 2003). In recent decades, collocation analysis has been applied to investigate syntactic and semantic phenomena in Western languages (Boogaart et al. 2014; Church et al. 1991; Gries and Stefanowitsch 2004; Mukherjee and Gries 2009; Stefanowitsch and Gries 2003, 2008). For instance, Church et al. (1991) investigated the differences in meaning between *strong* and *powerful* by looking at the words in associations with them; Gries and Stefanowitsch (2004) compared the ditransitive construction and the *to*-dative construction by analysing the collocates of these two constructions at the verb slot. Similar studies are available in Chinese, where quantitative methods and tools are employed in the field of computational linguistics. For instance, Huang et al. (2005) and Huang et al. (2015) explored

the possibilities for extracting grammatical classes from corpora with the *Sketch Engine* platform (Kilgarriff et al. 2014), and Gong et al. (2007) applied a frequency-based collocational approach to search for lexical mappings in Chinese based on a large-scale corpus. In addition, Sun et al. 孙茂松等 (1997) explored statistical measures to evaluate collocations in Chinese by their association strength and dispersion. Using statistical tests, You and Wang 由丽萍, 王素格 (2005) examined the distributions and rules of verb collocations and tried to define proper search windows in analysing different types of verbal structures.

However, collocation analysis has not gained much attention in discourse studies, especially those on the Chinese language. The majority of studies provide qualitative analyses and often refer to anecdotal examples to illustrate their claims. Of course, this qualitative approach is valuable in itself because the categories for classification must first be identified before they can be counted (McEnery and Wilson 2001; Schmied 1993). Other studies do provide quantitative data but are restricted in the sense that they present only percentages or frequencies. Without inferential statistics, it is difficult to generalize the conclusions beyond actual observations (McEnery and Hardie 2012; Núñez 2007).

For instance, in analyses of discourse relations and connectives as discourse markers, the main methods used are introspective studies with individual examples (e.g., Deng 邓雨辉 2007; Guo 郭继懋 2008; Zhao 赵新 2003). Deng 邓雨辉 (2007) claimed that the two result connectives 因而 *yin'er* 'as a result' and 因此 *yinci* 'as a result', despite both expressing the meaning of 'as a result', differ in their syntactic positions and the types of discourse segments they connect: *yinci* often appears after the subject of the subsequent clause and can be used to connect larger discourse units such as paragraphs, while *yin'er* is more restricted to connections between clauses and sentences and is used mostly at the initial position of the subsequent clause before the subject. Zhao 赵新(2003) reported a tendency of the connective 于是 *yushi* 'that's why' and 从而 *cong'er* 'thereby' to be used to connect actions instead of states, while *yinci* 'as a result' is more often used in sentences that describe states. Some studies provide corpus-based analysis with frequency counts (e.g., Li 李晋霞 2011; Li and Liu 李晋霞, 刘云 2004; Xing 邢福义 2002). For example, Li and Liu 李晋霞, 刘云 (2004) revealed that certain connectives, such as 既然 *jiran* 'since', tend to co-occur with discourse clauses expressing subjective opinions. Xing 邢福义 (2002) conducted a small-scale corpus-based study on novels and texts on political theory, finding that 由于 *youyu* 'because' is used mainly in argumentative texts and seldom in narrative texts.

Li et al. (2013) and Li et al. (2016) took the analysis of Chinese connectives as discourse markers a step further by investigating their use in different genres (argumentative, informative and narrative) and by applying inferential statistics to the data. They proposed an operational model with four dimensions to analyze the subjectivity of sentences marked by connectives: the domain of relations (epistemic, volitional content and nonvolitional content), propositional attitude (speech act/judgement, mental fact and physical fact), the presence of *SoC* (*Subject of Consciousness*, an illocutionary agent who is the speaker responsible for the reasoning), and the identity of *SoC* (the author, current speaker and character types). With a series of regression analyses on the four dimensions, Li et al. (2013) have shown that *yushi* 'that's why' and *yin'er* 'as a result' differ in the volitionality they express in causal relations. *Yushi* is used mainly to express volitional content relations and prefers contexts with intentional physical/mental acts and explicit SoC (example (1a); *yin'er*, in comparison, prefers nonvolitional content relations and no SoC (example (1b)).

(1) Examples of *yushi* 'that's why' and *yin'er* 'as a result'

a. 当地的经济危机已经持续一段时间, **于是**李明决定去国外申请工作。
Dangdi__de__jingji__weiji__yijing__chixu__yi__duan__shijian,
yushi__Li__Ming__jueding__qu__guowai__shenqing__gongzuo.
Local__MOD__economy__crisis__already__last__one__CL__time,
*CONJ*__NAME__decide__go__abroad__apply__job.
The local economy has been gloomy for a while, that's why Li Ming decides to apply for jobs abroad.

b. 当地的经济危机已经持续一段时间, **因而**失业率持高不下。
Dangdi__de__jingji__weiji__yijing__chixu__yi__duan__shijian,
yin'er__shiye__lü__chigao__buxia.
Local__MOD__economy__crisis__already__last__one__CL__time,
*CONJ*__unemployment__rate__keep:high__NEG:down.
The local economy has been gloomy for a while, as a result the unemployment rate stays at a high level.

In addition, Li et al. (2013) have shown that certain connectives display a profile that is robust across informative, narrative and argumentative genres, whereas other connectives appear to be genre-sensitive. The differences between the two connectives *yushi* and *yin'er* in terms of volitionality, for instance, remain salient across different genres.

The analysis of Li et al. (2013) illustrated the usage patterns of connectives based on manually annotated categories and a restricted sample of data. If the model and dimensions defined in their study are robust, we can expect to find converging evidence from different measures and in a larger-scale dataset. Statistical collocation analysis of the contexts of the two connectives may reveal collocation patterns that correspond to the differences between *yushi* and *yin'er* in various dimensions. For instance, the presence of an illocutionary agent and volitionality in the context presented in the model should be contextual features of *yushi* instead of *yin'er*. Such contextual feature differences can be captured well by collocation analyses. An attractive option for studying the use of discourse connectives from a more comprehensive view is to investigate discourse connectives in relation to discourse features and other discourse elements. Studying a word in its context provides more insights into the properties of the word, as Firth (1957: 11) argued: "you shall know a word by the company it keeps". From this perspective, collocation analysis based on associations between words is considered a suitable choice.

The purpose of the paper, therefore, is to provide an overview of methodological issues and solutions in the practice of performing a Chinese collocation analysis and to illustrate the value of collocation analysis for discourse studies with an example study on two Chinese connectives, which have been claimed to be different in terms of volitionality. In Section 2, we will discuss parameters that define the notion of collocation and introduce statistical methods to assess whether words in the context of a target word should be considered a collocate. Within the framework that defines and evaluates collocations, we discuss practical choices to make when applying collocation analysis in Chinese discourse studies in Section 3, such as word segmentation, the definition of search span, and the selection of corpora. Alongside the methodological discussions, we introduce a case study in Section 4 to exemplify the application of collocation analysis. We investigate *yushi* and *yin'er*, which are two synonymous result connectives but have been claimed to express different types of causal relations in discourse. The research questions of the case study come from both a theoretical perspective and a methodological perspective: how do contextual features of the two connectives reflect their properties in terms of encoding volitionality in causal relations? Do results from large-scale statistical collocation analysis converge with the previous findings from manual corpus-based analyses on a comparatively limited scale, i.e., is *yushi* more volitional than *yin'er*?

2 Review of collocation analysis

To conduct a collocation study, one must make decisions regarding a variety of dimensions or parameters. In a way, they thereby create their definition of the notion of collocation. Section 2.1 reviews five parameters that determine the type of elements under investigation according to the framework of Gries (2013). Section 2.2 illustrates different ways to determine the frequency at which these elements have to co-occur before they are considered a collocate, as well as practical decisions to make regarding the choice of measures.

2.1 Definition of collocations

Researchers first have to select "the nature of the elements" to be observed (Gries 2013: 138). Originally, the notion of collocation was introduced for characteristic and frequently recurring word combinations (Firth 1957). This focus on words is also apparent in Evert's (2008: 1214) definition – "a combination of two words that exhibit a tendency to occur near each other in natural language, i.e. to co-occur". Evert (2008) also noted, however, that a restriction on the word level is not necessary: the concept of collocation and the methodology can be applied to the co-occurrences of linguistic units, including morphemes, phrases and constructions.

The second and third parameters can broaden or restrict the type of elements that are considered collocates. The second parameter addresses the degree of lexical and syntactic flexibility of the collocates involved (Gries 2013). For instance, in the case of words, researchers may be interested in co-occurrence with exactly the same form –, e.g., looking at collocates of the noun *woman* – or they may increase the flexibility of their approach by focusing on lemmas (e.g., by including both *woman* and *women* as inputs in their collocation analysis). The third parameter concerns the role that semantic unity and semantic non-compositionality or non-predictability play in the definition; often, it is assumed that the elements considered as collocates exhibit something unpredictable in terms of form and/or function (Gries 2013).

A fourth parameter concerns the number of collocates that make up the collocation (Gries 2013: 138). In most cases, this value is "two", but the number of collocates is not restricted to this value. An N-gram analysis, for example, allows collocations composed of a sequence of N words in a fixed order, which could result in bigrams (N = 2), trigrams (N = 3), etc., depending on the value of N that is chosen (De Kok and Brouwer 2011).

The fifth parameter is the distance and/or (un)interruptability of the collocates (Gries 2013: 139). The most frequently used option is to focus on elements that are directly adjacent. Alternatively, researchers may be interested in elements that are syntactically or phrasally related but not necessarily adjacent, or they can investigate collocates that are more distant but still co-occur within a window of N words or within a specific unit, such as a sentence.

In sum, by making choices in line with the parameters distinguished by Gries (2013), researchers can develop their own definitions of the collocations that feature in their research. Awareness of these parameters help researchers to link their research questions with what the method enables them to do and facilitate interpretation of the output.

2.2 Applying statistics to determine collocations

In addition to the parameters that define the type of elements that are examined in a collocation study, Gries (2013) also distinguishes a parameter that concerns the frequency of the elements under investigation: one needs to decide upon the number of times an expression must be observed before it is counted as a collocate.

Some previous Chinese studies investigated the meaning of linguistic elements by looking at the expressions they co-occur with (e.g., Wang 王灿龙 2006; Yin 尹洪波 2011; Zhang 张焕香 2011). Some calculate raw frequencies of the co-occurrences of expressions in texts (Qi

齐春红 2007; Tang and Zhu 唐钰明, 朱玉宾 2008; Yin 尹洪波 2011), as is also common in the case of N-gram analyses. Although these studies have offered appealing accounts for linguistic phenomena in Chinese, inferential statistics are necessary to establish generalizable conclusions based on observed occurrences. Researchers could start by looking at collocates that occur more frequently than expected by chance, but there are thresholds and statistical scores other than raw frequencies of co-occurrence to measure the associations between target elements and collocates (e.g., PMI, Dice, Delta P, Odds Ratio, Chi-square, G^2, as illustrated below). By calculating additional statistics for collocates, one can rank the relevant collocates and set the threshold above a certain value as the cut-off line for "important" collocates or select the top N items (e.g., top 50, 100, etc.).

The statistical examination involves measures for the association between target words (nodes) and candidate collocates. Despite the variations of different collocation types, the measures for the association are all derived from the same contingency table: Table 1 (adopted from Gries (2013: 140), which is comparable to Evert's (2008) contingency table). Every word pair is referred to as $word_1$ (target word) and $word_2$ (candidate collocate), and *a*, *b*, *c*, *d* are the observed co-occurrence frequencies of the respective combinations. Expected frequencies of occurrences (a', b', c', d'), which are the occurrences of each combination under the null hypothesis that $word_1$ and $word_2$ are independent of each other (Evert 2005), can be calculated on the basis of *a*, *b*, *c*, *d.* The observed frequencies and the corresponding expected frequencies are used to calculate the strength of association between the target word ($word_1$) and each of the particular candidate collocates ($word_2$). The advantage of this method is as follows: the observed frequency of a word pair (*a*) is never evaluated in isolation but rather with regards to reference levels (*b*, *c*, *d*), which produces association scores that are robust for words of different frequencies and corpora of various sizes.

	Word2: present	Word2: absent	Totals
Word1: present	*a*	*b*	*a+b*
Word1: absent	*c*	*d*	*c+d*
Totals	*a+c*	*b+d*	*a+b+c+d*

Table 1. Co-occurrence table

(adopted from Gries 2013: 140)

The strength of associations between word pairs can be evaluated by association coefficients, which are from two types of measures (Evert 2008; Speelman 2021): effect size measures (measuring coefficients such as PMI, Dice, Delta P, and Odds Ratio) and statistical significance measures (coefficients such as Chi-squared (χ^2), log-likelihood measure (G^2), t, and Fisher). Effect size measures evaluate the magnitude of the difference between the observed co-occurrences and expected co-occurrences (Evert 2008; Gries 2013). For details on the formulas applied by these measures, see Evert (2008, Section 4.2) and Speelman (2021, Section 3.2.2). Association scores produced by effect size measures indicate how strong the attraction or repulsion is between the target word and the collocate. A brief summary of the interpretations of association scores from different measures is listed in Table 2.

Effect size measures	Attraction	Repulsion	Neutral
PMI	>0	<0	0
Odds Ratio	>1	<1	1
Delta P	[0, 1]	[-1, 0]	0
Dice	Approaching 1	n.a.	n.a.

Table 2. Summary of some effect size measures
(based on Evert 2008; Speelman 2021)

Different from effect size measures, significance measures evaluate the difference between the observed co-occurrences and expected co-occurrences from the perspective of a statistical test: how much evidence is there to establish an actual difference? Statistical association measures based on the amount of evidence include tests such as the Chi-squared (χ^2) test, log-likelihood (G^2) test, t-test, z-test, and Fisher test. Details on the calculation of association scores in these tests have been illustrated by Evert (2008, Section 5.2) and Speelman (2021, Section 3.2.3). With a low p-value (usually < .05) from these tests, an attraction between a collocate and the target word is established when the observed frequencies are significantly higher than the expected frequencies; repulsion is at stake when observed frequencies are lower than the expected frequencies.

The choice of the 'right' measure for a study is often open to debate. In practice, the PMI measure provides the most straightforward association coefficient to interpret – the log score of the ratio between the observed co-occurrence and the expected co-occurrences. Delta P seems to be a wise choice if one wants to make a psycholinguistic account for data since it has received more experimental support (Gries 2013). Despite the insights that effect size measures can bring, one downside of these measures is that they are unreliable with low-frequency data, which is due to their mathematical property of using 'direct estimates that do not take sampling variations into account' (Evert 2008: 1237). Significance measures, on the other hand, calculate collocation strengths based on the amount of evidence. However, according to Evert, some of them may suffer from the problem of either overestimating significance (such as the Chi-squared test and z-test) or underestimating significance (such as the t-test). Theoretical and technical accounts for the differences among various association measures have been discussed extensively by Evert (2005, 2008), Gries (2013), Gries and Stefanowitsch (2004), Pecina (2009) and Wiechmann (2008).

To obtain collocation results that not only have an acceptable effect size but also are supported by a sufficient amount of evidence, it is generally advisable to include both results from effect size measures and those from significance measures in a collocation report. One preferred way is to have a list of collocates ranked by a significance measure and one ranked by an effect size measure and then take the collocates in the overlap of the two lists as important collocates. Alternatively, one can use a test from one of the two types as the major measure (i.e., a log-likelihood, G^2) to produce a list of important collocates and apply a secondary criterion (a threshold of association score or top-N) on the basis of a test from the other type of measure (i.e., PMI).

3 Collocation analysis for Chinese discourse studies

In this section, we discuss practical decisions to make when conducting collocation analysis for Chinese discourse studies, in line with the framework of the parameters introduced in Section 2. Subtopics include the definition of collocates in Chinese (Section 3.1) and a suitable search span (target context) for discourse studies (Section 3.2). Apart from these parameters, we also propose that collocation analysis in the domain of discourse studies take genre issues

seriously because linguistic phenomena targeted by discourse studies are often sensitive to genre differences. Thus, the selection of corpora/sub-corpora for analysis is also important (Section 3.3).

3.1 Defining collocates in Chinese

Three of the parameters for collocation analysis (Gries 2013, Section 2.1) relate to the definition of meaningful linguistic units that can be considered collocates – possible candidates include morphemes, words, lemmas, etc. For the Chinese language, there is a long-running debate about whether the basic unit of Chinese is word or character (Pan 潘文国 2002; Xu 徐通锵 1994; Zhao 赵元任 1975). In natural language processing, word-based models and character-based models have been extensively tested and compared, and both receive credibility in text classifications and part-of-speech tagging (Gao et al. 2003; Ng and Low 2004; Zhang et al. 2003). To conduct discourse studies in Chinese with collocation analysis, however, we have two reasons to target words instead of characters. First, because the meaning of a character in Modern Chinese may differ depending on the word it is embedded in (Chang et al. 2008; Lee et al. 2014), a character-type collocate would be ambiguous in meaning. Thus, the results of the collocation analysis would be less interpretable if characters are the target collocates. Second, discourse studies usually adopt a global perspective, for instance, to explore contextual features in the scope of a discourse segment (e.g., clauses, sentences and paragraphs) and relations between segments, instead of addressing specific questions on the distributions of particular characters in texts. Therefore, a character-based approach would lead to less clear interpretations without bringing extra theoretical benefits for a discourse study.

To investigate the collocation patterns of words, we first need to define what a word is. Many Western languages use white spaces to separate words. Differences in spacing, however, may result in different outcomes. For instance, in English, *football player* is an expression composed of two individual words, while its Dutch counterpart *voetbalspeler* is written as a compound without a space between the component words. In the search for the collocates of the target *coach* in texts with the words *football/voetbal*, *player/speler* and *football player/voetbalspeler*, the English words *football* and *player* will appear only as two separate collocates; their combination will not appear in the analyses. In a Dutch search, *voetbal*, *speler* and *voetbalspeler* all appear as collocates. This example illustrates that word segmentation matters for the identification of target words and their collocates.

Unlike most Western languages, the Chinese writing system does not use white spaces to separate words, which makes the identification of words an issue. To apply collocation measures on the association of Chinese words and obtain reliable results, the first step is to correctly identify word boundaries. An economical choice is to use a well-segmented corpus for analyses, such as *Lancaster Corpus of Mandarin Chinese* (McEnery and Xiao 2003) and the *UCLA Written Chinese Corpus* (Tao and Xiao 2012). However, researchers may sometimes have to work with corpora that are not segmented for some practical reasons, for example, when a well-segmented corpus cannot provide sufficient data that are available only in a large-scale raw corpus. In this case, applying automatic word segmentation tools for Chinese can be an alternative solution. The recent development of natural language processing has contributed to the field with a large variety of segmentation tools (Liu and Wei 2008; Li and Guo 2016; Long et al. 龙树全等 2009; Wang and Guan 王晓龙，关毅 2005), and some major tools are listed in Appendix 1.

The method of using word segmentation tools facilitates further collocation analysis, but it is subject to two limitations. First, the accuracy rates of segmentation tools vary across text types, but none of the tools is 100% accurate. Therefore, we cannot expect the segmentation tools to produce perfectly annotated/segmented texts. Second, by performing collocation analysis with segmented texts, one must accept the definition of words used by the

segmentation system by default. That is, one can be forced to select words that are defined as 'words' according to the given segmentation system. If certain elements are ignored because they are not identified by the segmentation system, certain collocates may go unnoticed in the collocation study.

Therefore, it is important to choose a word segmentation tool that suits the target corpus and the research question a study aims to explore. First, the selected segmentation tool should have a high tested accuracy rate on the chosen corpus or corpora of similar types. For instance, if the research is performed with a narrative corpus, it may be wise to use segmentation tools with high accuracy rates in fiction, such as the *Stanford word segmenter* (Tseng et al. 2005). For research on argumentative texts, the *LTP cloud*[1], which has high accuracy rates tested in *People's Daily* newspaper data, would be a preferred choice. Segmentation systems that are based on large-scale dictionaries and well tested in balanced corpora, such as *NLPIR-ICTCLAS*[2], can be a suitable choice for research on texts from various sources.

3.2 Meaningful search span at the discourse level

The next question for collocation studies is: what is the context within which the collocates of the target word are detected (cf. the fifth parameter – distance – and/or (un)interruptability of the collocates discussed in Section 2.1)? One of the approaches is to set an arbitrary size of the search span, for example, five words to the left and five to the right of the target word. The words frequently occurring within that span size are considered to be collocates. This intuitive approach disregards meaningful discourse boundaries, thereby increasing unexpected noise in the data. For example, the border of the context may be put in the middle of a long sentence or clause, which creates a loss of data in comparison to an analysis in which the entire sentence is used as the span size. Similarly, sentences shorter than the set span size generate extra collocates from the preceding or following contexts, which also affects the calculation of association strengths.

For discourse studies, we suggest adopting a span size that makes sense at the discourse level, namely, a discourse segment such as a sentence, a clause, a paragraph or even a whole document, instead of a context containing an arbitrary number of adjacent words. A follow-up question is: what can be counted as a discourse segment? Discourse segments have been defined as chunks of text expressing a common purpose (Grosz and Sidner 1986) or a common meaning (Hobbs et al. 1988). Two general routines have been used to define the minimal unit of discourse: sentences (Hobbs et al. 1988) and clauses, as is common in the cognitive approach to coherence relations (Sanders et al. 1992) and annotations based on the rhetorical structure theory (Carlson and Marcu 2001; Mann and Thompson 1988). A practical concern in this area is related to the speciality of the Chinese discourse structure. The sentence boundaries in Chinese discourse are not as strict as those in Western languages. Example (2), taken from the CCL corpus (Zhan et al. 2003), gives a brief idea of what a Chinese 'sentence' could look like.

(2) 由于中期报告所载明的内容涉及到公司最基本的情况， 关系到广大投资者的权益， **所以**， 股票或者公司债券上市交易的公司在依法制定中期报告后， 应当依法将中期报告提交给国务院证券监督管理机构和证券交易所， 以使上述机构加强对上市交易的股票或者公司债券的监管， 保护广大投资者的合法权益。(Zhan et al. 2003)[3]

Since the content of the interim report concerns the basic situation of the company, concerns the benefit of many investors, ***so****, companies which issue public-traded stocks*

[1] https://www.ltp-cloud.com/intro. Accessed 26 May 2019.
[2] http://ictclas.nlpir.org/. Accessed 26 May 2019.
[3] Pinyin and gloss translations are not included in this particular example because of space limitations.

> *and corporate bonds, after they have made the interim report according to the law, should submit the reports to the securities regulatory body of the State Council and the Stock Exchange, in order to ensure the supervision and regulation of such institutes on the public-traded stokes and bonds, to protect the legal rights of investors.*

The whole paragraph in (2), with this vast amount of information, is presented as one long sentence (marked by one full stop) in on-going Chinese text. If the whole sentence is taken as one single discourse segment in a collocation analysis in which the researcher is interested in collocates within a more local context, the results would be too noisy. For instance, for the case study aiming at investigating the contextual features of two connectives – 于是 *yushi* 'that's why' and 因而 *yin'er* 'as a result' – the research question concerns whether the clauses connected by the connectives are physical facts or judgements, whether a volitional mind is involved in the clause, and so forth. Therefore, it is more reasonable to include one complete clause before and one after the connective as the context instead of having an arbitrary number of words with an inevitable loss of potential important elements or having long sentences as the context with an inclusion of extra irrelevant information.

Annotated corpora, such as *HIT IR-Lab Chinese Dependency Treebank*, offer neat and manifest segmentation systems and provide tags and unique identifiers for sentences/clauses, which can be efficient for detecting sentence/clause boundaries. However, applying the system of the corpus means that one accepts the segmentation of the corpus and disregards more flexible approaches. Some raw corpora, such as the CCL corpus, by contrast, do not provide any segmentations or annotations. For these corpora, the segmentation tools discussed in Section 3.1 can be applied to obtain similar tags for punctuations as one can get from an annotated corpus. For example, the *NLPIR-ICTCLAS* and the *Corpus Word Parser* mark the punctuations as */w* in the annotated texts they offer. The */w* marks commas or full stops, and can be applied to identify the boundary of clauses and/or sentences. *NLPIR-ICTCLAS* provides a more detailed repertoire of tags (e.g. */ww* for a question mark, */wj* for a full stop, */wf* for a semicolon, etc.), with which individual and flexible decisions can be made on the boundaries of discourse segments.

One important note is that one can never achieve perfection in the detection of discourse segments because with an automatic coding system, there are always cases that would have been segmented in another way in manual coding. The advantage of using raw corpora along with segmentation tools is that researchers have the freedom and flexibility to define discourse segments as they wish.

3.3 Selection of the corpus

For discourse studies, text properties such as genres, channels, and registers are important. If a study aims to address questions that are specific to certain genres or modalities, it is necessary to select a representative corpus or a sub-corpus from a balanced corpus. As Biber (1993) argued, prior identification of the genre categories can guarantee a good representation at the genre level. The genres, channels, and registers of a corpus matter for collocation analysis in at least two ways. First, researchers can add credence to the findings by checking whether collocation patterns are consistent across genres, channels, registers, or sub-corpora (Gries 2013). Stefanowitsch and Gries (2008) found some constructions to be more channel sensitive than others: the active construction exhibits sensitivity to differences related to spoken vs. written channels, while the passive construction consistently shows construction-specific preferences for certain types of collocates regardless of the channel types.

Second, researchers may formulate hypotheses about the effects of genres/channels/registers on collocation patterns. In other words, genre, channel, or register

types can be treated as an independent variable influencing the choice of words/word forms as the collocates of the target word (Gries and Stefanowitsch 2004). For example, in their study of Chinese causal connectives, Li et al. (2013) examined sentences containing connectives from three different genres: news reports, opinion pieces, and novels. The first two genres were taken from *People's Daily Online*, and the narrative genre was from the CCL corpus. Li et al. (2013) hypothesized that genre has an impact on the degree of subjectivity of a text and might therefore affect the meaning and use of Chinese causal connectives. For example, opinion pieces typically express the writer's point of view and aim to convince the reader by presenting arguments and are therefore likely to display an overall higher degree of subjectivity than news reports that are more descriptive and informative in nature. The results of this corpus-based analysis reveal that three connectives (i.e., *kejian* 'so', *yin'er* 'as a result', and *yushi* 'so/therefore') display robust profiles, whereas the distribution of two other connectives (*suoyi* 'so/therefore' and *yinci* 'so/therefore') are subject to genre changes (cp. Li et al. (2016) for a similar analysis on reason connectives *jiran* 'since', *yinwei* 'because' and *youyu* 'as').

Corpus size and accessibility also need to be taken into account when performing a collocation analysis. Despite the advantages of segmented and/or well-annotated corpora, many annotated corpora are limited in corpus size, which restricts the number of occurrences of words. Some association measures, however, are vulnerable to low frequencies of occurrences of items. For instance, the log-likelihood measure (G^2) test always requires a minimum frequency of three for statistical stability, and the p-values provided by a Chi-squared (χ^2) test would become unreliable if one of the expected values in the contingency table is less than five (Speelman 2021). To maintain statistical reliability, larger corpora are sometimes preferred for collocation studies.

The accessibility of texts also matters for collocation analysis. With full texts available, association scores can be automatically calculated based on the total number of words in the corpus and the observed frequencies by analytic tools such as *AntConc*, *WordSmith* and *R* packages (for instance, *mclm* by Speelman 2021). For corpora without full texts available, information on corpus size (in terms of the number of words/characters) is necessary to allow estimations on the occurrences of words based on the total size of the corpus.

Some corpora provide possibilities for on-site collocation analysis, such as the *Academia Sinica Balanced Corpus of Modern Chinese* (Chen et al. 1996), the *Lancaster Corpus of Mandarin Chinese* (McEnery and Xiao 2003), the *UCLA Written Chinese Corpus* (Tao and Xiao 2012), and the online corpus software interface *Sketch Engine* (Kilgarriff et al. 2014). With online inquiries on the corpus, for instance, top N collocates can be listed and ranked by the values of MI, MI3, Dice, t, z and log-likelihood. The information on collocation behaviours of words provided by these corpora can be widely applied to lexicography, language teaching, and linguistic research.

An overview of the available Modern Chinese corpora can be found in Appendix 2. The appendix provides information on the possibilities and characteristics of each corpus, such as the corpus size, segmentation information and availability of texts.

4 Case study: distinctive collocation analysis of two causal connectives

To exemplify the collocation analysis method to solve discourse-related questions, we conducted a collocation study to explore the differences between two connectives,于是 *yushi* 'that's why' and 因而 *yin'er* 'as a result', and the discourse relations they express. Recall that

in a discourse study performed with manual corpus-based analyses by Li et al. (2013), the two causal connectives differ in the following dimensions:

(i) *Yushi* is used mainly to express volitional content relations, which can be paraphrased as "P leads to intentional physical act/mental act that Q", while *yin'er* has a preference for non-volitional content relations – "P leads to the physical fact/mental fact that Q, and no intention is involved in Q" (Li et al. 2013: 87).

(ii) In terms of the presence of SoC, *yushi* has a preference for explicit and character-type SoC (*Subject of Consciousness*, Section 1), whereas *yin'er* prefers no SoC for the relation.

In the current study, we examined the differences between the two connectives and the relations they express by studying their contexts in a large-scale corpus. We performed *distinctive collocation analysis* (see Section 4.1 for details) to automatically extract the contextual information of the two connectives. Contextual features of the two connectives have been examined to address the following questions:

(1) How do contextual features of the two connectives reflect their properties in terms of encoding volitionality in discourse relations, i.e., what are the type of relations they express and the type of SoC they pattern with?

(2) Do results from large-scale statistical collocation analysis converge with the previous findings from manual corpus-based studies on a comparatively small scale, i.e., is *yushi* more volitional than *yin'er*?

4.1 Method

We performed a series of distinctive collocation analyses on the distribution of linguistic elements around the two connectives in comparison with each other. The analyses were conducted using the software *R* (R Core Team 2018) with the R package *mclm_0.1* (Speelman 2021). Distinctive collocation analysis (Church et al. 1991), as a type of collocation analysis, can be used to detect potential differences between comparable words (e.g., *yushi* 'that's why' and *yin'er* 'as a result' in example (1)). The critical difference between a distinctive collocation analysis and a simple collocation analysis is that the former takes the context of the competitor word as the reference context instead of the rest of the corpus. For example, two separate simple collocation analyses for the connectives *yushi* and *yin'er* would provide two lists of collocates, one for each of the two connectives, given the words occurring in the entire corpus, whereas a distinctive collocation analysis would generate one list of collocates that prefer the context of one connective over the other. This analysis can be applied to words as well as constructions (see the distinctive collexeme analysis in Gries and Stefanowitsch 2004, 2010).

The dataset was obtained from the CCL corpus (Zhan et al. 2003), which is a large collection of raw texts from different genres and modalities. CCL offers free, unlimited download of items. In total, 25,143 sentences containing *yin'er* and 59,358 sentences with *yushi* were archived. We downloaded a maximal number of 300 characters in the preceding context of the connectives and another 300 characters in the following contexts to obtain contexts large enough for different analyses. The preceding and following contexts did not exceed paragraph breaks.

In line with Gries's (2013) parameters in defining collocations, we first made the decision regarding the nature of elements; in this case, collocates were words instead of characters because Chinese characters are often polysemous, and it is difficult to determine their meanings on their own (see discussion in Section 3.1). The raw texts from CCL were segmented and

tagged with *NLPIR-ICTCLAS* segmenter (*NLPIR python wrapper*[4]). The *NLPIR-ICTCLAS* segmenter added spaces around each word and attached POS tags to words and punctuations. With the words segmented, we can obtain collocation results as the meaningful words. In terms of the parameter of lexical flexibility, we decided to investigate words of exactly the same form instead of lemmas. The number of words that make up a collocation pattern was two, one target word and one collocate word.

For the distance of the collocate, the search span, we included one clause before and one clause after the connectives. Commas, full stops, question marks, exclamation marks, semicolons and ellipses were chosen as the clause delimiters. Distinctive collocation analyses were performed to compare the contextual features of *yushi* and *yin'er*, with the context of *yushi* as the target context and the context of *yin'er* as the reference.

The association strengths of collocations were measured by G^2 and PMI. Collocates of *yushi* in reference to *yin'er* were ranked by G^2, which does not specify the direction of the association. Therefore, the top collocates ranked by G^2 included words that were strongly attracted to the *yushi* context in comparison to the *yin'er* context, as well as words that were strongly repelled by the *yushi* context in comparison to the *yin'er* context. The former type of words was considered the collocates of *yushi*, and the latter type of words were the collocates of *yin'er*. The top 30 words ranked by G^2 were considered important collocates of either *yushi* or *yin'er*. A secondary criterion, PMI, was applied: words in attraction to *yushi* and repulsion to *yin'er* all had a PMI above 0, and words in attraction to *yin'er* and repulsion to *yushi* had a PMI below 0. The PMI score also indicates the effect size of the distinctiveness of a word.

In addition to the analysis of all texts archived from CCL, we also conducted a genre-specific collocation study. In the analysis presented in Section 4.3, we split the raw dataset into two types according to the source of texts: narratives (e.g., fiction, biographies, stories, etc.) and non-narratives (e.g., newspapers, instructions, scientific theses, etc.). With this division, we can compare the collocations of *yushi* and *yin'er* in different genres. In the analysis for each genre type, the contexts of *yushi* and *yin'er* in both genres included one clause before and one clause after the connectives.

4.2 Results and discussion on general collocation patterns

In this distinctive collocation analysis (results presented in Table 3), the strong collocates of either of the two connectives were ranked by G^2 score, which was used as the main measure of association strength. The top 30 words that have the highest G^2 scores stand out as important collocates. Since G^2 only indicates the distinctiveness without specifying the direction of collocation, the collocates list in Table 3 includes both the collocates of *yushi* 'that's why' and those of *yin'er* 'as a result'. The PMI scores, as the secondary criterion, indicate the direction of collocation (PMI above 0: collocates of *yushi*; PMI below 0: collocates of *yin'er*) and the effect sizes of the association.

[4] https://github.com/haobibo/ICTCLAS_Python_Wrapper. Accessed 24 January 2017.

Rank		Collocates of *yushi*	Frequency (obs. vs exp.)	G^2	PMI
1	了	*le ASP(PFV)*	(18866: 14886)	3196.95	0.34
2	他	*ta 'he/him'*	(10260: 7549)	3042.44	0.44
3	我	*wo 'I/me'*	(5839: 4077)	2496.00	0.52
4	她	*ta 'she/her'*	(4170: 2882)	1907.59	0.53
5	便	*bian 'thereupon'*	(3244: 2190)	1734.18	0.57
7	就	*jiu 'just'/'then'*	(7101: 5588)	1215.78	0.35
8	去	*qu 'go'*	(2504: 1767)	990.86	0.50
12	又	*you 'further'*	(3579: 2743)	765.63	0.38
13	一	*yi 'one'*	(7449: 6184)	746.54	0.27
16	说	*shuo 'say'*	(2340: 1758)	587.27	0.41
19	想	*xiang 'think'*	(1298: 911)	534.66	0.51
20	那	*na 'that'*	(1505: 1086)	510.24	0.47
21	把	*ba (disposal construction)*	(2698: 2102)	501.22	0.36
22	来	*lai 'come'*	(2634: 2059)	475.82	0.36
23	到	*dao 'arrive'*	(3501: 2828)	466.49	0.31
24	着	*zhe ASP(IPFV)*	(2362: 1846)	428.06	0.36
25	走	*zou 'walk'*	(981: 685)	415.83	0.52
26	开始	kaishi 'begin'	(1101: 783)	414.20	0.49
29	你	*ni 'you'(singular)*	(1077: 773)	377.84	0.48
Rank		**Collocates of *yin'er***	**Frequency (obs. vs exp.)**	**G^2**	**PMI**
6	由于	*youyu 'since'*	(294: 1042)	1424.57	-1.83
9	的	*de (particle)*	(36410: 40129)	973.09	-0.14
10	是	shi (BE) 'is/are'	(5616: 7084)	798.50	-0.33
11	具有	*juyou 'have'*	(139: 533)	777.48	-1.94
14	经济	*jingji 'economy'*	(288: 722)	679.00	-1.33
15	和	*he 'and'*	(3602: 4691)	659.81	-0.38
17	发展	*fazhan 'develop'*	(374: 790)	564.46	-1.08
18	较	jiao 'compare'	(119: 405)	535.13	-1.77
27	受到	shoudao 'suffer'	(150: 402)	411.07	-1.42
28	对	dui 'towards'	(2114: 2762)	395.28	-0.39
30	社会	*shehui 'society'*	(397: 716)	365.59	-0.85

Table 3. Top 30 collocates ranked by connective and G^2

From the collocation list in Table 3, we observed the following tendencies of collocations in the *yushi* context:

(i) Aspect markers that are attached to verbs in Chinese preferred the context of *yushi* compared to the context of *yin'er*, such as *le* ASP (PFV) and *zhe* ASP (IPFV);

(ii) The cognition verb expressing mental acts *xiang* 'think' and the communication verb reflecting mental acts *shuo* 'say' both appeared more frequently in the context of *yushi*;

(iii) There were more motion verbs such as *dao* 'arrive', *zou* 'walk', and *lai* 'come' in the *yushi* context;

(iv) Adverbials such as *bian 'thereupon'* and *jiu 'just'/'then'*, which are associated with the expressions of actions, co-occurred more often with *yushi*;

(v) *Ba* (disposal construction), as an element of the disposal verbal construction, was also strongly related to the *yushi* context instead of *yin'er* context;

(vi) Pronouns that refer to an animate person collocated more with *yushi*, such as *ta* 'he/him', *ta* 'she/her', *wo* 'I/me' and *ni* 'you'(singular).

Contextual features of *yin'er* also surfaced in the collocation analysis:

(i) The stative verbs *juyou* 'have' expressing procession and the copula *shi* 'is/are' were used to describe states co-occurring more often with *yin'er*;
(ii) *Shoudao* 'suffer', as an element of passive constructions that involve less volitionality than the active form of verbs, appeared as a collocate of *yin'er*;
(iii) *Jiao* 'compare', which is used in comparison structures, stood out as the collocate of *yin'er*.

The results from the collocation study are consistent with Li et al.'s (2013) findings in a corpus-based analysis of *yushi* and *yin'er* and the introspective study of Chinese connectives by Zhao 赵新 (2003). As Li et al. (2013) have suggested, *yushi* is more associated with volitional content relations than *yin'er* is. Volitional content relations involve more intentional physical acts/mental acts and are more likely to have an explicit and character-type SoC that is responsible for the reasoning or action in the causal relations. Therefore, it can be predicted from their corpus-based study that in the context of *yushi*, there might be more words related to actions and motions than in the context of *yin'er*, as well as more expressions referring to an illocutionary agent. In our collocation analysis, these predictions are borne out. The collocation results showed a tendency to have more "acts" and volitionality in the context of *yushi*: more verbs, aspect markers, adverbials and verbal constructions were found for the context of *yushi* compared to *yin'er*. Moreover, there were more pronouns in the context of *yushi* compared to the context of *yin'er*, including the first-person, second-person and third-person pronouns, which could serve as the illocutionary agent (the character-type SoC) to perform intentional physical acts/mental acts.

By contrast, *yin'er* has a preference for non-volitional content relations in the study by Li et al. (2013): it patterns with both physical/mental facts that involve no volition and prefers no SoC in the context. Therefore, *yin'er* was expected to appear more in a context containing statements and descriptions of physical facts than *yushi*, and without an illocutionary agent. The cooccurrence patterns of *yin'er* and its collocates support the predictions. The context of *yin'er* contained more verbs describing states instead of actions (e.g., *you* 'have' and the copula *shi* 'is/are'); the verb *shoudao* 'suffer', which is an important element of Chinese passive constructions, co-occurred more often *yin'er*. Collocates of *yin'er* also included *jiao* 'compare', which constitutes a structure to describe comparisons of states between objects and events. These contextual features of *yin'er* lend credence to the non-volitionality encoded by *yin'er* as a connective for non-volitional content relations indicated by Li et al.'s corpus-based study.

In summary, the results of the collocation analysis, with the inclusion of all instances of the two connectives in a large-scale corpus, have supported the findings from the discourse analysis based on a relatively restricted sample of data. The converging results illustrate the validity and value of collocation methods in investigating Chinese discourse.

4.3 Results and discussion on collocations in different genres

Genre differences are critical to discourse studies and may influence collocation patterns. Narrative genres are assumed to have high volitionality in the texts, while argumentative and informative genres are supposed to be less volitional. Will the general features of a genre overwrite or smooth the distinctiveness of some words in the context of one connective over the other? To address this question, we performed a set of distinctive collocation analyses for *yushi* and *yin'er* per genre type (narrative and non-narrative). Table 4 shows the top 30 collocates of *yushi* in narratives and non-narratives, and Table 5 shows the top 30 collocates of *yin'er* in the two types of genres.

Collocates	Narratives			Non-narratives		
	Frequency (obs. vs exp.)	G^2	PMI	Frequency (obs. vs exp.)	G^2	PMI
Pronouns						
我 wo 'I/me'	4099: 3837	160.08	0.10	1740: 1030	1062.96	0.76
他 ta 'he/him'	7043: 6703	149.11	0.07	3217: 2137	1136.40	0.59
她 ta 'she/her'	3406: 3196	122.09	0.09	764: 452	468.15	0.76
他们 tamen 'they'	-	-	-	1281: 969	203.46	0.40
Verbs/aspect markers						
了 le ASP(PFV)	10332: 9746	315.04	0.08	8534: 6199	1819.31	0.46
去 qu 'go'	1701: 1545	156.73	0.14	803: 514	339.71	0.64
走 zou 'walk'	751: 670	107.71	0.17	-	-	-
说 shuo 'say'	1709: 1579	100.07	0.11	-	-	-
来 lai 'come'	1539: 1429	76.85	0.11	1095: 811	201.56	0.43
着 zhe ASP(IPFV)	1610: 1507	63.29	0.10	-	-	-
起来 qilai 'get up'	671: 610	59.02	0.14	-	-	-
把 ba (disposal construction)	1579: 1482	55.74	0.09	1119: 815	230.53	0.46
向 xiang 'lead'	654: 598	49.84	0.13	627: 428	190.25	0.55
下 xia 'go down'	727: 668	49.75	0.12	-	-	-
决定 jueding 'decide'	496: 449	49.55	0.14	-	-	-
个 ge 'singe'	1091: 1017	48.76	0.10	-	-	-
开始 kaishi 'begin'	-	-	-	575: 344	333.67	0.74
到 dao 'arrive'	-	-	-	1627: 1213	285.98	0.42
想 xiang 'think'	-	-	-	458: 270	285.38	0.76
发现 faxian 'discover'	-	-	-	525: 333	233.47	0.66
找 zhao 'find'	-	-	-	231: 125	208.18	0.89
Adverbials						
便 bian 'thereupon'	1678: 1507	204.06	0.16	1566: 871	1258.61	0.85
就 jiu 'just'/'then'	3803: 3539	179.06	0.10	3298: 2395	696.13	0.46
又 you 'further'	1975: 1837	93.60	0.10	1604: 1119	432.36	0.52
Other						
一 yi 'one'	4028: 3810	109.15	0.08	3421: 2712	374.80	0.34
地 di (particle)	1927: 1822	52.63	0.08	-	-	-
两 liang 'two'	685: 628	49.70	0.13	-	-	-
纷纷 fenfen 'numerously'	-	-	-	219: 121	180.13	0.86

Table 4. Collocates of *yushi* in the top 30 collocates list per genre[5]

The major categories of the contextual features of *yushi* remained robust in both narrative and non-narrative genres, as Table 4 illustrates. Collocates of *yushi* included pronouns that indicate illocutionary agents and verbs and related elements such as aspect markers and adverbials that represent actions and volitions. Although the specific words that surfaced from different genres were not exactly the same, the major features of the context were stable across narratives and non-narratives.

[5] The italicized collocates are also in the *top 30 collocates list* for both genres in Table 3 (Section 4.2). Empty slots of a word in one type of genre mean that the word did not surface in the top 30 collocates of the connective *yushi* in this genre.

Collocates	Narratives			Non-narratives		
	Frequency (obs. vs exp.)	**G^2**	**PMI**	**Frequency (obs. vs exp.)**	**G^2**	**PMI**
由于 youyu 'since'	65: 142	199.75	-1.13	229: 742	778.68	-1.70
的 de particle	16272: 16791	122.65	-0.05	20138: 22126	382.04	-0.14
受到 shoudao 'suffer'	48: 86	83.27	-0.84	102: 268	221.08	-1.40
使 shi 'make' (causative verb)	377: 453	77.78	-0.27	-	-	-
是 shi (BE) 'is/are'	2864: 3041	71.97	-0.09	2752: 3862	650.58	-0.49
对 dui 'towards'	993: 1091	58.10	-0.14	1121: 1560	250.12	-0.48
所 suo 'that which'	306: 361	51.97	-0.24	-	-	-
更 geng 'more'	375: 435	51.84	-0.21	-	-	-
具有 juyou 'have'	29: 51	48.09	-0.82	110: 392	450.14	-1.83
和 he 'and'	-	-	-	2091: 2807	370.30	-0.42
经济 jingji 'economy'	-	-	-	256: 548	326.00	-1.10
较 jiao 'compare'	-	-	-	98: 304	303.73	-1.63
发展 fazhan 'develop'	-	-	-	331: 597	244.95	-0.85
它 ta 'it' (inanimate)	-	-	-	517: 812	218.57	-0.65

Table 5. Collocates of *yin'er* in the top 30 collocates list per genre[6]

As exhibited in Table 5 for collocates of *yin'er*, in both narratives and non-narratives, the verb *shoudao* 'suffer', as a part of passive construction, the copula *shi* (BE) 'is/are' and the state verb *juyou* 'have' expressing procession appeared as important collocates of *yin'er*, which demonstrated the non-volitionality of contexts around *yin'er*. Some noticeable elements stood out only for the narratives: *geng* 'more' is used in comparison; *shi* 'make' (causative verb) appears as part of the construction *A shi B+verb* 'A makes/forces B to do/to be…', in which case, it is not B's intention or volition to do/to be so. These collocates added evidence regarding the non-volitional nature of the *yin'er* context. For non-narratives, we found collocates *jiao* 'compare', as part of comparative construction and *fazhan* 'develop', which is a verb describing the change of state of an inanimate agent such as a company or an organization. Unlike other pronouns, in non-narratives the pronoun *ta* 'it' appeared as the collocate of *yin'er* instead of *yushi*. An important feature of *ta* 'it' is that it refers only to inanimate agents, which are not subject to volitions. Therefore, in both narratives and non-narratives, the context of *yin'er* exhibits a clear property of non-volitional features in contrast to that of *yushi*.

Taking together the results from Tables 4 and 5 and comparing them with those in Table 3, we found a moderate overlap between the collocates list in narratives with that generated by collocation analyses on both types of genres in Table 3 (66.7%, 20 out of 30). The major contextual features of *yushi* regarding volitionality reported in Table 3 were also found as distinctive features of the context of *yushi* in Table 4. Therefore, the general features of narrative genres did not smooth out the distinctive features of the *yushi* context compared to the *yin'er* context. In other words, even in the narrative genre, which has an overwhelming use of verbs and verbal constructions, the connective *yushi* still attracted more active verbs and verbal constructions that are closely related to volitionality than the connective *yin'er* did.

The overlap of collocates in non-narratives and those in Table 3 is relatively higher (80%, 24 out of 30). Most of the distinctive features remained for the non-narrative genres, with

[6] The italicized collocates are also in the *top 30 collocates list* for both genres in Table 3 (Section 4.2). Empty slots of a word in one type of genre mean that the word does not surface among the top 30 collocates of the connective *yin'er* in this genre.

several new words boosted by the genre feature of non-narratives, such as *faxian* 'discover' and *fenfen* 'numerously'; importantly, this did not change the general collocation patterns.

In summary, although different genres have generated variations in the exact rankings of words in the collocation lists, they did not overwrite the main categories of collocates in terms of volitionality in discourse relations. This finding is also consistent with Li et al.'s study (2013), which argued for the stability of the features of the two connectives across different genres.

5 Conclusion

The current paper presented a complete procedure to perform collocation analysis and a series of methodological considerations, with a case study as an exemplification. We have demonstrated the need to segment *words* as collocates in Chinese and the advantage of analyzing collocations within a meaningful discourse unit. We have also shown in a genre-specific analysis how genres might influence the distribution of collocates. The output of the collocation analysis, as an explorative method based on large-scale corpora applying statistical measures, is in line with the results from the manual analysis on limited numbers of samples (Li et al. 2013). Collocation patterns have shown a clear tendency to have more expressions of volitionality in the context of *yushi*, which has previously been claimed to be a connective typically used for volitional content relations. *Yin'er*, by contrast, appeared in contexts with more elements related to non-volitional relations, which is also consistent with prior studies.

Although the collocation method, as an explorative approach, could be limited in the sense that it cannot provide decisive conclusions on the exact factors influencing collocations, it provides an efficient tool to analyze discourse. The merit of the method is that it allows intuitive interpretations of clusters of collocates and inferential statistics of word frequencies, which measures attraction or repulsion between words/expressions in a more stable and reliable way than analysis based on anecdotal examples. The example study, along with the methodological discussions in the paper, demonstrate the advantages of collocation analysis to address discourse-related questions, and we expect future discussions and practice to cover more theoretical and technical details on the discourse analysis of Chinese, as well as other languages.

Appendix 1 Overview of segmentation tools for Chinese

1. *NLPIR-ICTCLAS* (*Institute of Computing Technology, Chinese Lexical Analysis System*)[7] is an HHMM-based (Hierarchical Hidden Markov Model) framework for Chinese word segmentation and annotation. HHMM-based frameworks allow analysis from five levels: atom segmentation, simple and recursive unknown words recognition, class-based segmentation and POS (Part of Speech) tagging (Zhang et al. 2003). *NLPIR-ICTCLAS* provides word segmentation with an accuracy rate of 98.5% (see the evaluation on NLPIR-ICTCLAS-3 in Feng and Zheng 奉国和, 郑伟 2011).
2. *Hailiang intellectual word segmentation*[8] – *research version* is a well-developed word segmentation system for Chinese. It has applied adequate algorithms to better settle the ambiguous segmentations and unknown word recognition. The reported accuracy rate of *Hailiang* in segmenting the closed corpus CCL (Center for Chinese Linguistics of Peking University) is 99.6% (Feng and Zheng 奉国和, 郑伟 2011).
3. The *Stanford word segmenter* (Tseng et al. 2005) is a Java implementation of the CRF-based (Conditional Random Field) Chinese word segmenter. It has achieved a high F-score in four Mandarin Chinese corpora: 0.947 for Academia Sinica Corpus, 0.943 for the corpus by University of Hong Kong, 0.950 for the corpus by Peking University, and 0.964 for the corpus by Microsoft Research Asia (Tseng et al. 2005).
4. The *LTP-cloud* (*Language Technology Platform*)[9] is an on-line system which provides word segmentation, part-of-speech tagging, syntactic analysis and annotations of semantic roles. Among those functions, the word segmentation module has achieved high accuracy in People's Daily newspaper data (development set: Precision = 0.973, Recall = 0.972, F-score = 0.973[10]; test set: Precision = 0.972, Recall = 0.970, F-score = 0.972).
5. *SCWS* (*Simple Chinese Word Segmentation*)[11] is an open source word segmentation engine based on inserted dictionaries. The tested accuracy is 95% and the recall is 91%.
6. Other tools: *ChineseTA*[12], *Corpus Word Parser*[13], *Pan Gu Segment*[14], *MMSEG* system[15] and *Jieba* Chinese text segmentation tool[16].

[7] http://ictclas.nlpir.org/. Accessed 20 June, 2020.
[8] http://bigdata.hylanda.com/smartCenter2018/index. Accessed 20 June, 2020.
[9] http://www.ltp-cloud.com/. Accessed 20 June, 2020.
[10] *Precision* and *recall* are evaluation measures widely used in machine learning, natural language processing and information retrieval, etc. *Precision* equals the total number of relevant items retrieved divided by the total number of items that are retrieved; *recall* is the total number of relevant items retrieved divided by the total number of relevant items in the database (Ting 2010). *F-score* is the weighted harmonic mean of *Precision* and *Recall*.
[11] http://www.xunsearch.com/scws/. Accessed 20 June, 2020.
[12] http://www.svlanguage.com/. Accessed 20 June, 2020.
[13] http://www.cncorpus.org/. Accessed 28 August, 2016.
[14] https://archive.codeplex.com/?p=pangusegment. Accessed 28 August, 2016.
[15] http://technology.chtsai.org/mmseg/. Accessed 20 June, 2020.
[16] https://github.com/fxsjy/jieba. Accessed 20 June, 2020.

Appendix 2 Overview of Modern Chinese written corpora

Name	Size	Content	Period	Segmentation	Text availability
Academia Sinica Balanced Corpus of Modern Chinese 4.0 (现代汉语平衡语料库)	sentences: 1,396,133; word tokens: 11,245,932; word types: 239,598	Balanced: written (report, review, biography, diary, poem, letter, etc.), oral (scenario, conversation, speech, conference); Style: narration; argumentation, exposition; description; Medium: newspaper, general magazine, academic journal, textbook, thesis, audio/visual medium, conversation/interview, etc.; Topics: philosophy, natural science, social sciences, arts, general/leisure, literature	since 1996	yes	full text available web search, no download limit
BCC Chinese corpus (北京语言大学现代汉语语料库)	characters: 1,500,000,000	Balanced: newswire, web language, literature, technology, etc.	released in Sep 2014	yes	web search, download max. 1000/ 10,000 items; randomization allowed
Chinese Gigaword Fifth Edition	n.a.	Newswire	1990-2010	no	full text (available from Linguistic Data Consortium, Catalog No. LDC2016T13)
Chinese POS Tagged Corpus	words tokens: 5,000,000	Balanced: Newswire, fictions, proses, scripts, descriptive documents, letters, argumentative texts, biographies, conversations, essays.	developed 2002.10-2003.10	yes	full text (available from CLDC, No. CLDC-LAC-2003-003)
Chinese Treebank 9.0	sentences: 132,076; word tokens: 2,084,387; characters: 3,247,331	Newswire, magazine articles and government documents, chat messages and transcribed conversational telephone speech.	1994-2006	yes	full text (available from LDC, Catalog No. LDC2011T13)
Cncorpus (国家语委现代汉语通用平衡语料库)	characters: 19,455,328; word tokens: 12,842,116 (incl. punctuations)	Balanced: Textbooks; Humanities and social science (history, economics, literature, arts, etc.); science (agriculture, engineering and technology, etc.), newspapers and magazines, practical writing documents (official documents, letters, advertisements, etc.)	1919-2002	yes	full text (raw corpus available from Chinese Linguistic Data Consortium(中文语言资源联盟) No. CLDC-LAC-2006-001) web search, download max. 5000 items
HIT-CIR Chinese Dependency Treebank	without relations: 50,000 sentences; with relations: 10,000 sentences	Newswire	n.a.	yes	full text

Name	Size	Content	Period	Segmentation	Text availability
Lancaster Corpus of Mandarin Chinese version 2 (LCMCv2)	word tokens: 1,000,000 (incl. punctuation)	Balanced: Press (reportage, editorials, reviews); Religion; Skills, trades and hobbies; Popular lore; Biographies and essays; Miscellaneous (reports and official documents); Science: academic prose; General fiction; Mystery and detective fiction; Science fiction; Adventure and martial arts fiction; Romantic fiction; Humor	1991 (+/- 2 years)	yes	full text (available from the Oxford Text Archive-Catalogue No. 2474) web search, no download limit; randomization allowed
LIVAC (Linguistic Variations in Chinese Speech Communities) synchronous corpus	2.7 billion characters	Newswire	1995-2018	Partially segmented (680 million)	web search
National broadcast Media Language corpus	characters: 241,316,530 (incl. punctuations)	Broadcast. Media: video, broadcast; Mode: conversations, monologues, dialogues and general; register: conversations, announcements, tutorials.	2008-2013	yes	web search, no download limit
Peking University CCL Corpus	characters: 581,794,456	Balanced: fictions, newspaper, conferences, translated literature, blogs, etc.	n.a.	no	web search, no download limit
Sketch engine - 9 Chinese sub-corpora	9 sub-corpora with a total collection of 4,099,628,033 tokens	Newswire, web texts, and parallel corpora of web texts.	Depends on sub-corpora	yes	web search/full text, download limit depending on sub-corpora
The Corpus of Chinese Compound Sentences (汉语复句语料库)	compound sentences: 658,447; characters: 44,395,000	Newswire	n.a.	no	web search, no download limit
The Corpus of Contemporary Novels (当代小说语料库)	sentences: 657,136	Fictions.	n.a.	no	web search, no download limit
The UCLA Written Chinese Corpus (2nd edition)	word tokens: 1,119,930	Same as LCMC	2000-2012	yes	web search, no download limit, randomization allowed
Tsinghua Chine se Treebank	characters: 1,000,000	Literature (fictions, proses, scripts), media (biographies, news), academic works, practical texts.	developed 1998 – 2003	yes	full text (raw corpus available from CLDC, No. CLDC-LAC-2003-005

Acknowledgements

Part of this paper is based on Chapter 2 of the first author's Ph.D. thesis (Wei 2018), which was enabled by the Chinese Scholarship Council (grant number 2013077220042, 2013).

References

Biber, Douglas. 1993. Representativeness in corpus design. *Literature and linguistic computing* 8(4): 243–257.

Boogaart, Ronny, Timothy Colleman, and Gijsbert Rutten. 2014. Constructions all the way everywhere: Four new directions in constructionist Research. In *Extending the Scope of Construction Grammar*, ed. Ronny Boogaart, Timothy Colleman, and Gijsbert Rutten, 1–14. Berlin: Mouton de Gruyter.

Carlson, Lynn and Daniel Marcu. 2001. *Discourse tagging reference manual* (ISI technical report. ISI-TR-545). Online: http://www.isi.edu/~marcu/discourse/.

Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, ed. B.-S. Park and J.B. Kim, 167–176. Seoul: Kyung Hee University.

Chang, Pi-Chuan, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. *Proceedings of the 3rd Workshop on Statistical Machine Translation*, Columbus, Ohio, 224–232.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical acquisition: Exploiting on–line resources to build up a lexicon*, ed. Uri Zernik, 115–164. New Jersey: Lawrence Erlbaum Associates.

Church, Kenneth and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1), 22–29.

De Kok, Daniël and Harm Brouwer. 2011. *Natural language processing for the working programmer*. Online: http://freecomputerbooks.com/books/nlpwp.pdf.

Deng, Yuhui. 邓雨辉. 2007. Analysis of the usage of *yinci* and *yin'er* 果标“因此”和“因而”的用法辨析. *Journal of Guangzhou University (Social Science Edition)* 广州大学学报（社会科学版）6(8), 79–82.

Evert, Stefan. 2005. *The statistics of word cooccurrences: Word pairs and collocations.* Ph.D. dissertation. University of Stuttgart.

Evert, Stefan. 2008. Corpora and collocations. In *Corpus linguistics: An international handbook*, ed. Anke Lüdeling and Merja Kytö, 1212–1248. Berlin: Mouton de Gruyter.

Feng, Guohe, and Wei Zheng 奉国和, 郑伟. 2011. Review of Chinese automatic word segmentation 国内中文自动分词技术研究综述. *Library and Information Service* 图书情报工作 55(2): 41–45.

Firth, John. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in linguistic analysis*, ed. John Firth, 1–32. Oxford: Blackwell.

Gao, Jianfeng, Mu Li, and Chang-Ning Huang. 2003. Improved source-channel models for Chinese word segmentation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, 272–279.

Gong, Shu-Ping, Kathleen Ahrens, and Chu-Ren Huang. 2007. Chinese Sketch Engine and mapping principles: A corpus-based study of conceptual metaphors using the building source domain. *Proceedings of the 8th Chinese Lexical Semantics Workshop*, Hong Kong, 130–136.

Gries, Stefan. 2013. 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics* 18(1): 137–165.

Gries, Stefan, and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1): 97–129.

Gries, Stefan and Anatol Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In *Empirical and experimental methods in cognitive/functional research*, ed. Sally Rice, and John Newman, 73–90. Stanford, CA: CSLI Publications.

Grosz, Barbara and Candace Sidner. 1986. Attention, intentions, and the structure of discourse. *Journal Computational Linguistics* 12(3): 175–204.

Guo, Jimao 郭继懋. 2008. A contrastive analysis between sentences with *yinwei suoyi* and *jiran name* "因为所以"句和"既然那么"句的差异. *Chinese Language Learning* 汉语学习 3: 22–29.

Hobbs, Jerry, Mark Stickel, Paul Martin, and Douglas Edwards. 1988. Interpretation as abduction. *Proceedings of the 26th Annual Meeting, Association for Computational Linguistics*, New York, 95–103.

Huang, Chu-Ren, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychlý, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese Sketch Engine and the extraction of grammatical collocations. *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, Jeju Island, 48–55.

Huang, Chu-Ren, Jiafei Hong, Weiyun Ma, and Petr Šimon. 2015. From corpus to grammar: Automatic extraction of grammatical relations from annotated corpus. *Journal of Chinese Linguistics Monograph Series 25*, 192–221.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1(1): 7–36.

Lee, Chia-Ling, Ya-Ning Chang, Chao-Lin Liu, Chia-Ying Lee, and Jane Yung-jen Hsu. 2014. Semantic clustering of morphologically related Chinese words. *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Palo Alto, California, 3116–3117.

Li, Shihai and Shufen Guo. (2016). Collocation analysis tools for Chinese collocation studies. *Journal of Technology and Chinese Language Teaching* 7(1), 56–77.

Li, Fang, Jacqueline Evers-Vermeul, and Ted Sanders. 2013. Subjectivity and result marking in Mandarin: A corpus-based investigation. *Chinese Language and Discourse* 4(1): 74–119.

Li, Fang, Ted Sanders, and Jacqueline Evers-Vermeul. 2016. On the subjectivity of Mandarin reason connectives: Robust profiles or genre-sensitivity? In *Genre in language, discourse and cognition*, ed. Wilbert Spooren, Gerard Steen and Ninke Stukker, 15–49. Berlin/New York: Mouton de Gruyter.

Li, Jinxia 李晋霞. 2011. On the differences between *youyu* and *yinwei* 论"由于"与"因为"的差异. *Chinese Teaching in the World* 世界汉语教学 25(4): 490–496.

Li, Jinxia and Yun Liu 李晋霞, 刘云. 2004. The differences of *youyu* and *jiran* in subjectivity "由于"与"既然"的主观性差异. *Chinese Language* 中国语文 2: 123–128.

Liu, Jian and Cheng Wei 刘件, 魏程. 2008. Arithmetic research on Chinese segmentation 中文分词算法研究. *Microcomputer Applications* 微计算机应用 129(18): 11–16.

Long, Shuquan, Zhengwen Zhao, and Hua Tang 龙树全, 赵正文, 唐华. 2009. Overview on Chinese segmentation algorithm 中文分词算法概述. *Computer Knowledge and Technology* 电脑知识与技术 5(10): 2605–2607.

Mann, William and Sandra Thompson. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text* 8(3): 243–281.

Manning, Christopher and Hinrich Schütze. 2000. *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice.* New York: Cambridge University Press.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An introduction.* Edinburgh: Edinburgh University Press.

McEnery, Tony and Richard Xiao. 2003. The Lancaster Corpus of Mandarin Chinese. Paris / Oxford: European Language Resources Association / Oxford Text Archive.

Mukherjee, Joybrato and Stefan Gries. 2009. Collostructional nativisation in new Englishes: Verb-construction associations in the international corpus of English. *English World-Wide* 30(1): 27-51.

Núñez, Rafael. 2007. Inferential statistics in the context of empirical cognitive linguistics. In *Methods in cognitive linguistics*, ed. Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson, and Michael Spivey, 87–118. Amsterdam: John Benjamins.

Ng, Hwee Tou, & Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? Word-based or character-based? *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, 277–284.

Pan, Wenguo 潘文国. 2002. *Zibenwei yu hanyu yanjiu (Sinogram-based theory and Chinese studies)* 字本位与汉语研究. Shanghai: East China Normal University Press.

Pecina, Pavel. 2009. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1-2): 137–158.

Qi, Chun-hong 齐春红. 2007. On the co-occurrence rules of modal adverbs and sentence-ending auxiliary modal words 语气副词与句末语气助词的共现规律研究. *Journal of Yunan Normal University (Humanities and Social Sciences)* 云南师范大学学报（哲学社会科学版） 39(3): 125–130.

R Core Team. 2018. *R: A language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. http://www.R-project.org/. Accessed 27 February 2018.

Sanders, Ted, Wilbert Spooren, and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15(1): 1–35.

Schmied, Josef. 1993. Qualitative and quantitative research approaches to English relative constructions. In *Corpus based computational linguistics,* ed. Clive Souter, and Eric Atwell, 85–96. Amsterdam: Rodopi.

Speelman, Dirk. 2021. *Mastering corpus linguistics methods: A practical introduction with Antconc and R*. Hoboken, NJ: Wiley.

Stefanowitsch, Anatol and Stefan Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.

Stefanowitsch, Anatol and Stefan Gries. 2008. Channel and constructional meaning: A collostructional case study. In *Cognitive sociolinguistics*, ed. Gitte Kristiansen, and René Dirven, 129–152. Berlin: Mouton de Gruyter.

Sun, Maosong, Changning Huang, and Jie Fang 孙茂松,黄昌宁,方捷. 1997. Exploration of Chinese collocations with quantitative analysis 汉语搭配定量分析初探. *Chinese Language* 中国语文 1:29–38.

Tang, Yuming and Yubin Zhu 唐钰明, 朱玉宾. 2008. On the sentences with both passive construction and disposal construction 汉语被动/处置共现句略论. *Journal of Sun Yat-Sen University (Social Science Edition)* 中山大学学报: 社会科学版 48(1): 53–58.

Tao, Hongyin and Richard Xiao. 2012. *The UCLA Chinese corpus (2nd edition)*. UCREL, Lancaster University. https://www.lancaster.ac.uk/fass/projects/corpus/UCLA/. Accessed 27 February 2018.

Ting, Kai-Ming. 2010. Precision and recall. In *Encyclopedia of machine learning*, ed. Claude Sammut, and Geoffrey Webb, 781. New York: Springer US.

Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A Conditional random field word segmenter. *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*, Jeju Island, 168–171.

Wang, Canlong 王灿龙. 2006. On the co-occurrence of *mei (you)* and *le* 关于“没(有)”跟“了”共现的问题. *Chinese Teaching in the World* 世界汉语教学 1: 41–50.

Wang, Xiaolong and Yi Guan 王晓龙, 关毅. 2005. *Jisuanji ziran yuyan chuli (Natural language processing)* 计算机自然语言处理. Beijing: Tsinghua University Press.

Wei, Yipu. 2018. *Causal connectives and perspective markers in Chinese: The encoding and processing of subjectivity in discourse*. Utrecht: Utrecht University dissertation.

Wiechmann, Daniel. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4 (2): 253–290.

Xing, Fuyi 邢福义. 2002. On the semantic preference of the pattern introduced by *youyu* “由于”句的语义偏向辨. *Chinese Language* 中国语文 4: 337–342.

Xu, Tongqiang 徐通锵. 1994. Characters and Chinese syntactic structures“字”和汉语的句法结构. *Chinese Teaching in the World* 世界汉语教学 2: 1–9.

Yin, Hongbo 尹洪波. 2011. *Foudingci yu fuci gongxian de jufa yuyi yanjiu (Syntactic and semantic research on the co-occurrence of negatives and adverbs)* 否定词与副词共现的句法语义研究. Shanghai: Foreign Language Teaching and Research Press.

You, Lipingand Suge Wang 由丽萍, 王素格. 2005. Rules and distributions of Chinese verb-verb collocations 汉语动词-动词搭配规则与分布特征. *Computer Engineering and Applications* 计算机工程与应用 23:179–181.

Zhan, Weidong, Rui Guo, and Yirong Chen. 2003. *The CCL Corpus of Chinese Texts: 700 million Chinese characters, the 11th Century B.C. – Present*. Center for Chinese Linguistics of Peking University. http://ccl.pku.edu.cn:8080/ccl_corpus. Accessed 21 January 2017.

Zhang, Huanxiang 张焕香. 2011. Asymmetry of the word order in the collocation of frequency adverbs and negative adverbs 频度副词与否定副词共现时语序的不对称. *Journal of Capital Normal University (Social Sciences Edition)* 首都师范大学学报(社会科学版) 1: 113–118.

Zhang, Hua-Ping, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer NLPIR-ICTCLAS. *Proceeding of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, 184–187.

Zhao, Xin 赵新. 2003. Multi-dimension analysis on yinci, yushi and cong’er “因此、于是、从而”的多角度分析. *Linguistic Research* 语文研究 1: 26–29.

Zhao, Yuanren 赵元任. 1975. The concept, structure and rhythm of Chinese words 汉语词的概念及其结构和节奏, in *Zhao Yuanren Yuyanxue lunwen ji (Collection of Zhao Yuanren’s Linguistic Articles)* 赵元任语言学论文集 *(edition 2002)*. Beijing: The Commercial Press.