34

ONLINE LANGUAGE RESOURCES

Advances, applications and challenges

Weidong Zhan and Xiaojing Bai

Introduction

Alongside the dramatic development of the Internet and computer technology, the significance of language resources has become more and more evident in linguistic research and language teaching. The use of linguistic data has prevailed, unprecedentedly, in academia, industry and the government, following the new waves of artificial intelligence (AI) and natural language processing (NLP) brought by the rapid progress of mobile web, social media and machine learning, especially deep learning based on neural network models. Accordingly, a very large amount of Chinese language resources have been available in the past three decades (Zhan 詹卫东 2013a).

This chapter will introduce Chinese language resources online, including corpora, knowledge bases and related search engines, with extended discussions about their possible impact on linguistics and the challenges to their future development. This chapter will describe a survey of Chinese language resources currently online, introduce the application of these resources in linguistic research and language teaching and discuss the challenges to future development. The chapter will conclude by suggesting that there is a close connection between introspection-based theoretical analysis and data-driven statistical analysis.

Advances: a survey of Chinese language resources online

This survey was mainly conducted through web search engines (e.g. Google, Baidu, etc.) as well as resources collected by a few well-known scholars and teachers of Chinese[1] and related courses like Corpus Linguistics. Considering the co-existing simplified and traditional Chinese writing systems and the diverse resources in countries and regions where Chinese is not the official language, this survey covered (1) Mainland China; (2) Hong Kong and Taiwan; (3) other Asian countries (e.g. Japan, South Korea, Singapore, etc.); (4) North America; (5) Great Britain; and (6) the European continent.

Search queries included *corpus*, *knowledge base*, *lexicon*, *dictionary*, *Chinese* and *Mandarin*, with keywords highlighting the related applications, such as *natural language processing*, *machine translation* and *teaching Chinese as a second language*. Information on a great mix of resources was found, the web addresses of which are listed in the Appendix. Some

discussions on language resources for Chinese from different perspectives can be found in Chapter 31 by Huang and Xue (2019).

Three types of corpus-related resources, varying in corpus size, content and impact, were found, as follows: (1) web portals as repositories and distribution points for corpora, such as LDC and Chinese LDC; (2) online search engines equipped with large corpora, such as Word Sketch Engine, WebCorp Search Engine, and search engines with the CCL Corpus from Peking University, the BCC Corpus and HSK Writing Corpus from the Beijing Language and Culture University and the web corpora of Academia Sinica (Taiwan); and (3) smaller corpora developed by important academic institutions and researchers for the special purposes of Chinese teaching and research and Chinese information processing, such as the UCLA Written Chinese Corpus and the Hong Kong Bilingual Child Language Corpus from the Chinese University of Hong Kong.

Three types of knowledge bases, also termed lexicons or dictionaries, included (1) knowledge bases for academia and the information industry, such as the *Grammatical Knowledge-base of Contemporary Chinese*, the Chinese Semantic Dictionary and the Chinese Concept Dictionary from Peking University, Chinese WordNet and the Bilingual Ontological Wordnet from Academia Sinica and HowNet, which was built by MT researcher Zhendong Dong; (2) online lexicons, dictionaries and mobile applications available to the public for the transmission, teaching and use of Chinese, such as Souwen Jiezi from the Digital Library and Museum of the National Science Council (Taiwan) (http://words.sinica.edu.tw), the free Tradict.net for translators (Taiwan) (www.tradict.net/lang_guoyu.php) and 汉典Han Dian (Mainland China) (www.zdic.net/) for searches in vintage Chinese dictionaries, such as 康熙字典 *Kāngxī Zìdiǎn* and 说文解字 *Shuōwén Jiězi*; and (3) web portals or websites for knowledge bases of the world's languages, with Chinese and its dialects included, for the special purposes of cross-linguistic comparison, theoretical linguistic research, typological studies, etc., such as the resources page of the Association for Linguistic Typology (www.linguistic-typology.org/resources.html), the online World Atlas of Language Structures (WALS) (http://wals.info/), the UCLA Phonological Segment Inventory Database (UPSID), phonological data from the world's languages, which can be accessed via PHOIBLE Online (http://phoible.org/) and the P-base of sound patterns built by Dr Jeff Mielke at the University of Ottawa (http://aix1.uottawa.ca/~jmielke/pbase/).

Compared with the early development of modern corpora and linguistic knowledge engineering, the recent advances in the Chinese language resources mentioned above are impressive. Feng 冯志伟 (2006) traced how the corpora of English and Chinese developed since the 1960s and the 1980s, respectively. Early corpora mainly relied on manual work, such as examining small collections of sample texts used for counting the frequencies of characters and words. Today, a large quantity of multimedia data (e.g. textual, audio, etc.) from the Internet support a wide array of applications. Owing to the high cost of development and processing, some resources can only be accessed with granted permission and specific tools, but with the prevailing principle of open collaboration in the Internet era, many Chinese language resources are now accessible online.

Further, the ever-increasing storage and computational capacity of computers have made it possible to manage extremely large linguistic data sets, GB-sized or even TB-sized. For instance, the size of Chinese Web 5-gram Version 1 (https://catalog.ldc.upenn.edu/LDC2010T06; also accessible at https://archive.org/download/google_ngrams-chinese-simplified) released by LDC is approximately 30 GB. The data set contains Chinese word n-grams (from unigrams to 5-grams) and their frequency counts generated by Google Ngram[2] from Google's text corpora of books printed between the years 1500 and 2008.

In addition, the variety of corpora is greater than ever, and the designs of knowledge bases also vary. Some corpora have different annotations, such as segmented and POS-tagged, parsed as syntactic trees or with the semantic role labelled, semantic dependency marked and discourse relation glossed. Other corpora have different types of data, for instance: monolingual, bilingual, and multilingual; written and spoken; common expressions and Internet slang; child language and adult language; contemporary Chinese and Ancient Chinese; and the language of native speakers and second language learners. With knowledge bases, some are designed for general purposes, such as HowNet and the *Grammatical Knowledge-base of Contemporary Chinese*, while others are for special purposes, such as Emotion Ontology released by the Dalian University of Technology for sentiment analysis and Knowledge Base for the Qualia Structure of Nouns developed by Peking University for the semantic analysis of nouns and noun phrases (Yuan and Li 袁毓林, 李强 2014).

The rapid progress of Chinese language resources is, on the one hand, the result of ever-thriving corpora and data-driven empirical methods. On the other hand, consortiums like LDC have played a significant role in collecting and distributing resources. Founded in 1992, LDC published 164 Chinese data sets between 1994 and 2016, among which there are 109 text data sets, 54 audio data sets and one video data set. Some text data deals with the properties of words, phrases, sentences and discourses, while others are corpora built for machine translation and its evaluation, information retrieval, language modelling, etc. The text data mainly came from newspapers and magazines and the audio data mainly from telephone conversations, broadcast conversations, news broadcasts, etc. Unlike early corpora, which relied on printed materials, corpora nowadays collect more materials from sources like blogs, forums and newsgroups on the Internet. The annual statistics for LDC language resources are shown in Figure 34.1.

The advances in Chinese language resources are exciting and encouraging, but there are some reflections that may be worth considering. For instance, the development of Chinese

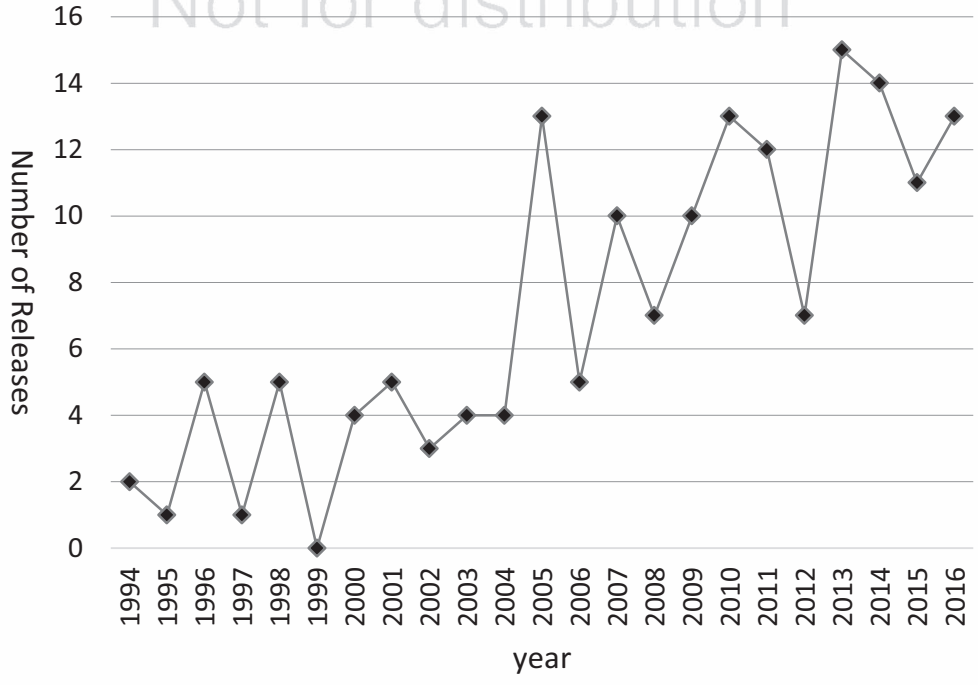


Figure 34.1 LDC Chinese language resources: annual releases (1994–2016)

language resources lags far behind that of English language resources in the sense of both volume and diversity. In 2016, LDC had 787 data sets in total, of which 434 data sets were for English. Moreover, among the 164 publications of Chinese resources at LDC, there are far more corpora (160 data sets) than knowledge bases or lexicons (four data sets), which mirrors the progress of Chinese language resources as a whole. The cost of constructing knowledge bases or lexicons is substantially high, considering the human input of refined linguistic knowledge. Large corpora, in contrast, can be collected and processed automatically, with less human input or even no expert knowledge, while more sophisticated annotations, particularly semantic annotations, have mainly been applied to corpora of a smaller size. Further, the potential of corpora for special purposes is great, such as interlanguage corpora, multimedia corpora and multimode corpora for teaching Chinese as a foreign language, discourse analysis, etc. Finally, while large corpora are much easier to compile today, the quality of these language resources is more likely to be overlooked. It has been reported, for example, that the Chinese data from Google Ngram between 1970 and 2000 is relatively more reliable than data from before 1940, with fewer OCR errors and hence a smaller amount of noise for automatic counts (http://digitalsinology.org/when-n-grams-go-bad/).

Applications

In the past few decades, corpus-based approaches have gained more and more ground. With big data and machine learning methods playing a decisive role in NLP today, the significance of language resources, and of corpora in particular, has been well acknowledged. The following discussion will focus on the roles of Chinese language resources in linguistic research and language teaching as well as in social progress. Due to limited space, only textual resources will be covered; audio and multimedia resources will not be included.

Using corpora in the study of synonyms

In teaching and learning Chinese as a first or second language, the clear understanding and the appropriate use of synonyms is stressed regularly. In this regard, corpus-based investigations have much to offer. While a native speaker may not be able to form consistent and convincing judgements on the distinction between synonyms, a corpus may provide sufficient samples, which, after being ranked and sorted, depict in detail how and why they differ.

Using a 500,000-character corpus of vernacular texts in the contemporary sense, Tao (2000) showed some systematic differences between eight synonymous adverbs that share the common meanings of absolute time and assertiveness. In the authoritative dictionary 现代汉语词典 *Xiàndài Hànyǔ Cídiǎn* (Modern Chinese Dictionary, Institute of Linguistics, CASS, 2011), these adverbs are deployed to define one another and are roughly grouped into three categories: (1) 从来 *cónglái* 'always', 向来 *xiànglái* 'always' and 一向 *yíxiàng* 'at all times'; (2) 根本 *gēnběn* 'at all', 本来 *běnlái* 'at first' and 全然 *quánrán* 'completely'; and (3) 一直 *yìzhí* 'always' and 始终 *shǐzhōng* 'all the time'. In addition, there are simple indications in some dictionaries, with no further explanation, about the usage of these adverbs, such as 从来 and 根本, which are mainly used in negative contexts. With quantitative evidence from the corpus, Tao (2000) investigated the syntactic environment, semantic prosody and textual linking functions of these adverbs and found more precise details about their distinctions. For instance, 从来, though usually followed by a negative element, is meant to negate an undesirable proposition or state-of-affairs in nearly all the examples retrieved from the corpus. In the following example (1), the speaker considers it undesirable to "get involved in lawsuits", but with the

help of 从来, the speaker expresses a positive stance. Such is not the case with 根本 when followed by a negative element, as can be seen in (2) where the speaker's stance is still negative:

1 银行有律师，靠我可打不赢官司，我**从来**不打官司。
Yínháng yǒu lǜshī, kào wǒ kě dǎbùyíng guānsi, wǒ cónglái bù dǎ guānsi.
'We have lawyers in the bank; they would never rely on me to win cases, and I never got involved in lawsuits'.

2 真的，我当时**根本**没想过会有今天这种事儿。
Zhēnde, wǒ dāngshí gēnběn méi xiǎng guo huì yǒu jīntiān zhè zhǒng shìr.
'Really, I never think about what is happening now'.

Other findings, such as 向来, 一向 and 全然, are not favoured in vernacular texts; 根本 appears in a context with a negative/unfavourable connotation, while 本来 has a neutral connotation, and 一直 requires a prior temporal/spatial phrase, while 始终 stands alone. Although Tao (2000) did not exhaust all the distinctions between the eight adverbs, which may be impossible even for a native speaker of Chinese, the results nevertheless shed some new light on such a complex issue that has mainly relied on the introspection of linguists and lexicographers for solutions.

Moreover, discourse context-based studies on linguistic properties have enabled a better understanding of language and in this regard the corpus approach has a part to play. As Tao (2000) pointed out, the traditional research of adverbs tended to limit the scope of investigation to isolated clauses/sentences. With the corpus, Tao provided evidence to argue that the appropriate use of many adverbs relies on a syntactic environment larger than a single clause/sentence. In (3), no grammatical problem can be found at the clause level, but the corpus data clearly shows that the adverb 一直 requires a temporal phrase, which is missing in this sentence and therefore renders it strange to native speakers. The corpus data also shows the common use of 根本 in conjunction with a repetition structure, as in (4), which might be overlooked when the adverb is observed in isolation:

3 我**一直**在看中文书。
Wǒ yìzhí zài kàn zhōngwénshū.
'I have been reading Chinese books'.

4 我没学过中文，**根本**没学过。
Wǒ méi xué guo zhōngwén, gēnběn méi xué guo.
'I didn't learn Chinese before. I never did'.

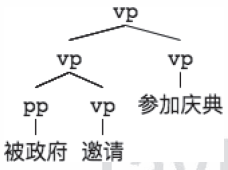
Tao (2000) presented an early application of corpora in studying Chinese synonyms. Larger corpora with segmentation and POS tags, which came later, made it possible to measure the pattern and the salience of collocations using statistics like Mutual Information, which indicates the mutual dependence between random variables (Huang et al. 2005; www.sketchengine.co.uk/documentation/statistics-used-in-sketch-engine/). Hong and Huang (2006) used collocating relations to present the difference between the two synonymous verbs 吃 *chī* 'eat' and 喝*hē* 'drink', based on the Chinese GigaWord Corpus of 1.1 billion Chinese characters. In addition, corpora of interlanguage, with incorrect word usages, have been found in recent studies on synonyms. Hong (2014), for instance, used a learner corpus of 300 million Chinese characters to investigate the usage errors of 方便 *fāngbiàn* 'convenient' and 便利 *biànlì*

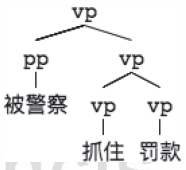
'convenient', which were compared with the proper usages of native speakers to show the different distributions of the synonyms.

Using treebanks in the analysis of syntactic ambiguities

A large corpus enables the observation of detailed language facts, as discussed previously. It may also facilitate the systematic understanding of a language as a whole (Zhan 詹卫东 2013b). In this section, the use of treebanks in the quantitative analysis of Chinese syntactic ambiguity will be discussed.

Studies on linguistic ambiguity in Chinese tend to focus on particular cases, such as 咬死了猎人的狗 *yǎo sǐ le lièrén de gǒu* 'bit and killed the hunter's dog or the dog that bit and killed the hunter'. The sequence (tagged as v-n-de-n) can form either a verb phrase or a noun phrase, leading to two readings and hence an "explicit ambiguity". It is, however, quite possible that a syntactically ambiguous sequence has only one reading, resulting in an "implicit ambiguity", such as the phrasal sequence pp-vp-vp in the following two examples:





被政府邀请参加庆典	被警察抓住罚款
bèi zhèngfǔ yāoqǐng cānjiā qìngdiǎn	bèi jǐngchá zhuāzhù fákuǎn
'be invited to attend a ceremony by government'	'be caught and fined by policeman'

For a systematic investigation into the possibility of linguistic ambiguity, treebanks can be used to extract context-free grammar rules (e.g., vp → pp vp, vp → vp vp), such as using a one-million-character treebank, for instance, to build a rule set with 11 phrase categories (tagged as ap, dp, np, pp, vp, etc.) to describe Chinese. Parsing the 1,331 sequences (e.g., np-vp-np), formed by any three of the tags, with the rules extracted from the treebank, the sequences are thereby categorized as (a) unacceptable, (b) unambiguous, (c1) explicitly ambiguous and (c2) implicitly ambiguous, as Zhan et al. 詹卫东等 (1999) proposed. The results are shown in Table 34.1 below, which quantitatively depicts the possibility of syntactic ambiguity in Chinese. For comparison, the corresponding statistics from Zhan et al. 詹卫东等 (1999) based on a manually defined rule set are also given in Table 34.1.

Among the acceptable sequences, there are far more ambiguous than unambiguous ones, which points to the underlying problems of phrase categorization in this case. The design of phrase categories needs to be optimized based on a better understanding of the language as a whole. This is an issue of importance, particularly when computers are used to assist in linguistic research.

Other linguistic studies have used treebanks to present the panoramas of language facts. For instance, Liu et al. (2009) used a dependency treebank to measure the dependency distance and direction, hence the complexity, of Chinese. The method was extended to measure the dependency distances across languages using an English and Czech dependency treebank (Jing and Liu 2015). Studies such as these are making substantial progress, with the support of sophisticatedly annotated corpora and greatly improved NLP technology.

Table 34.1 Probability of syntactic ambiguity calculated based on a manual rule set and a treebank rule set

Phrasal sequences \ *Rule sets*			*Manual rule set (246 rules)*		*Treebank rule set (657 rules)*	
			Count	*Percentage*	*Count*	*Percentage*
Total			$9^3 = 729$	100%	$11^3 = 1331$	100%
(a) Unacceptable			360	49.4%	661	49.66%
	(b) Unambiguous		84	11.5%	150	11.27%
	(c) Ambiguous	(c1) Explicit	194	26.6%	386	29.00%
Acceptable			Average	6.55%	Average	6.61%
		(c2) Implicit	91	12.5%	134	10.07%
			Average	2.37%	Average	2.29%

Using language resources in contrastive and typological analysis

In addition to corpora, knowledge bases are also widely used in linguistics, which will be the main focus of this section. As a database for linguistic typology, WALS currently contains data from 2,679 languages and facts about 192 grammatical features, each feature having between two and 28 different values. A huge number of facts about Chinese can be found there. Feng 冯胜利 (2015) made new arguments about the relation between tone, intonation and sentence-final particles: (1) all tonal languages have sentence-final particles; (2) non-tonal languages have no sentence-final particles (with exceptions); (3) more tones exist in a language with more sentence-final particles; (4) more sentence-final particles exist in a language with fewer intonations; and (5) tones in a language develop in parallel with its sentence-final particles. These arguments were further explored by Ye 叶述冕 (2016), with some affirmed and others negated. With the data from WALS, particularly the facts about tones and polar questions in 206 languages, Ye 叶述冕 (2016) investigated the typological correlations between tones, intonation and particles and argued against a salient typological correlation between tones and sentence-final particles. A correlation may exist, as non-tonal languages tend to have no sentence-final particles. However, tonal languages vary: those of Eurasia tend to have sentence-final particles, while those in America do not; and no tendency has been found with those in Africa and the Papuan-Austronesian region. Ye 叶述冕 (2016) further argued that there is a possible correlation between intonations and sentence-final particles yet pointed out the lack of language facts supporting the claim that "intonation and SFP are fundamentally two sides of the same coin" (Feng 冯胜利 2015).

Theoretical linguistic research tends to suggest and verify a hypothesis based on a small sample of language facts. With the increase in data from world languages, which are sophisticatedly annotated and easy to search, large-scale investigations across languages will better support linguistic typology and enable a deeper understanding of human languages, for which Ye 叶述冕 (2016) is a case in point. In fact, parallel corpora have played an important part in comparative linguistic research as well. Using a Chinese-English parallel corpus, Bai and Zhan 柏晓静，詹卫东 (2006) explored the differences between passive sentences (with the marker 被 *bèi*) in Chinese and those in English.[3] Similarly, Ma 马千 (2011) investigated the role of 这 *zhè* 'this' and 那 *nà* 'that' in definite marking and compared them with *the*, a similar expression in English.

Using corpora in sociolinguistic analysis

Between 2004 and 2008, China's Ministry of Education, in cooperation with universities and institutions, set up six national centres for the management and research of language resources from publications, broadcasts, the Internet, teaching materials, minority dialects and varieties of Chinese outside China, respectively (www.moe.gov.cn/s78/A19/A19_xglj/201309/t20130929_158028.html). This new initiative of managing linguistic data as public resources has been well received in academia and in society as well. Large dynamic corpora have thereby been released to support linguistic research and language teaching; on the other hand, they have provided solid evidence for the government and the public in learning the actual use of Chinese and its dialects in China today. Starting in 2006, the Report on Language in China has been published annually by the Department of Language Information Management at China's Ministry of Education, documenting the use of Chinese each year, informing the public about language services, facilitating government planning and promoting the healthy use of language in society (www.moe.edu.cn/s78/A19/yxs_left/moe_813/s237/; more details are provided by Li 李宇明(2007). Popular or new characters, words and expressions have been recorded, such as 互联网+ *hùliánwǎng*+ 'Internet Plus', 阅兵蓝 *yuèbīnglán* 'Parade Blue' and 重要的事情说三遍 *zhòngyào de shìqíng shuō sānbiàn* 'thrice to emphasize', which depict the kaleidoscopic quality of social life. Over the last 10 years, 5,514 new words have been collected in this annual report, such as 微博 *wēibó* 'microblog', 中国梦 *zhōngguómèng* 'China Dream', 微信 *wēixìn* 'WeChat' and 正能量 *zhèngnéngliàng* 'positive energy', highlighting the coming of each new trend in society (www.jyb.cn/china/gnxw/201510/t20151017_640001.html). The general public has also been invited to select the top 10 new words, catchphrases and Internet slang, which has greatly encouraged the public to participate in the ongoing development of language resources.

Moreover, the varieties of Chinese have been compared and contrasted by researchers in different countries and regions. The LIVAC Corpus (2.5 billion Chinese characters; www.livac.org) maintained by the City University of Hong Kong has collected texts from the newspapers of eight Chinese-speaking cities around the world. Such a corpus serves as a rich source of evidence for the variation of written Chinese in different regions. Zhan and Tao 詹卫东, 陶红印 (2016) built a corpus of written texts (approximately 100 million Chinese characters) from North America, with which words and their variations, together with their grammatical features, were investigated. The influence of English and the trace of southern dialects[4] of Chinese were found in the corpus data. There is, however, an obvious tendency for Chinese and its varieties to converge, owing to the ever-shortening distance between people in the net-connected global village. New usages in Chinese, carried by new media, have spread quickly around the globe, for which language resources are being developed to reflect more about the Chinese language's evolution.

Another example worth noting is the creation of a huge language resource to study "the global collective consciousness". The Global Database of Events, Language, and Tone (GDELT) contains "over a quarter-billion event records in over 300 categories covering the entire world from 1979 to present, along with a massive network diagram connecting every person, organization, location, theme and emotion" (www.gdeltproject.org/). The GDELT platform provides Global Content Analysis Measures to assess emotional undercurrents and reactions in articles in 15 languages, including Chinese. Such language resources capture ongoing stories around the world and thereby facilitate the study of human societal behaviours and beliefs.

Challenges: the future of Chinese language resources

In either theoretical studies (such as linguistics per se, linguistic typology, language acquisition, etc.) or applications (such as lexicography, teaching Chinese as a second language, language planning, etc.), Chinese language resources play active and important roles. Search tools for these resources have become more user-friendly and powerful, with various options for locating specific information. Tools for corpora like BCC and CCL, for instance, support advanced searches with patterns and wildcards.

It is worth noting, however, that the increasing size of linguistic data and the increasing cost and difficulty of developing them have caused new concerns about their quality and the validity of their support for research. When large amounts of texts in corpora come from the Internet, it is extremely hard to remove embedded errors and redundant data. Take WebCorp (www.webcorp.org.uk/live/index.jsp) as an example. This online tool connects to commercial search engines to retrieve hits for a user's query, takes the list of URLs returned by that search engine and extracts concordance lines from each of those pages to present examples of the user's query word or phrase in context. The concordance lines for the Chinese word sequence 虽然短暂 *suīrán duǎnzàn* 'though short' is shown in Figure 34.2, with substantial redundancy that is quite obvious.

Further, the achievements of Chinese language resources nowadays, if examined closely, are mainly the "low-hanging fruit", and there is still plenty. Real challenges remain in reaching the higher fruit. Chinese function words, for instance, have been a focus of attention in linguistic research and language teaching. Compared with synonymous content words, synonymous function words are far more difficult to set apart. Language resources for function words are inadequate, which also reflects the inadequate theoretical research and understanding in this regard. Sketch Difference has been used to make distinctions between content words (verbs like *chi* and *he* or adjectives like 方便 and 便利) (Hong and Huang 2006; Hong 2014). In the case of synonymous function words (一直 and 始终 or 根本 and 从来), however, Sketch Difference returns "insufficient data". Likewise, Word Sketch fails to find any collocations for 的 *de*, the most frequently used function word, though a simple query of the word in Sketch Search returns 99,838,775 hits. In another test with 了 *le*, Word Sketch returns collocations that confuse the word with its homograph 了 *liǎo*. To solve problems like these, more effort


```
……

10) http://www.le.com/tv/89028.html
Text, Wordlist, text/html, UTF8 (Content-type), 2017-01-01 (Copyright
footer)

29:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:25 第2集
30:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:24 第3集
31:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:25 第4集
32:…风起云涌的斗争，并由此讲述了他虽然短暂，但却充满了传奇的一生。 45:24 第5集
33:  风起云涌的斗争  并由此讲述了他虽然短暂  但却充满了传奇的  生   45:25 第6集
```

Figure 34.2 An example of duplicate sentences retrieved from WebCorp

in solving complex linguistic issues is required, which will lead to a better understanding of language, including its operation and development.

The challenge of Chinese language resources in the future also lies in the role that linguistic theories play in further exploiting these resources. Take the study of new senses of words as an example. The corpus approach helps greatly in collecting and presenting words in new contexts, such as the Chinese verb 刷 *shuā* 'brush', in the following collocations retrieved from the online CCL Corpus and BCC Corpus: 刷系统 *shuā xìtǒng* 'to update the operation system on the mobile phone', 刷微博 *shuā wēibó* 'to update the microblog very frequently', 刷街 *shuā jiē* 'to travel on the street with roller skates as a form of transportation or sport', 刷票 *shuā piào* 'to compete for a ticket online by frequently visiting the ticket-master's page', 刷人品 *shuā rénpǐn* 'to gain a good reputation by behaving well', etc. The meanings of the verb *shuā* in these new contexts, however, are not readily available in dictionaries; for instance, the verb has two senses recorded in the dictionary 现代汉语词典Xiàndài Hànyǔ Cídiǎn (Modern Chinese Dictionary, Institute of Linguistics, CASS 2011):

【动】 用刷子清除或涂抹 ：~牙|~鞋|~锅|用石灰浆~墙。
【Verb】 brush; scrub; clean or paint with a brush: brush one's teeth | brush shoes | clean (or scour) a pot | paint the wall with limewash; whitewash a wall
【动】 除名；淘汰："他不守劳动纪律，让厂里给~了|他报名参军，因视力不合格被~了下来。"
【Verb】 eliminate (through selection or competition); remove: 'He was dismissed from the factory due to his failure to follow the labour discipline. | He flunked the examination for army recruitment due to poor eyesight'.

In this case, corpus-based studies can be greatly assisted by the cognitive theory of conceptual blending. By "blending" the semantic components of the verb itself and those of the nominals in the new contexts, the meanings of the new usages can be better distinguished. For example, with the traditional usage 刷墙 *shuā qiáng* 'whitewash a wall', a new layer is added to the wall; with the new usage 刷系统 *shuā xìtǒng* 'to update the operation system on the mobile phone', similarly, new features are added to the mobile phone. The semantic description of 刷 in the new usage 刷系统, therefore, can be a "blend" – the semantic component of "adding something new" in the traditional scenario of the verb and the semantic component of "an electronic system with various features" in the new scenario of its collocate. Currently, it is easy for the corpus methods to retrieve evidence for words in new contexts, but it is too demanding for them to discover the underlying connections between semantic components and to define the new meanings automatically and systematically. In this regard, linguistic theories still have a leading role to play, so that the corpus evidence can be better exploited to make more sense.

Language resources today are mainly collections of language facts, the increasing volume of which does not mean an increasing amount of knowledge. Challenges remain in discovering and extracting linguistic knowledge from linguistic data automatically, but it is still a tempting and worthwhile endeavour.

Conclusion

The past three decades have witnessed the exciting progress of linguistic data in Chinese. Yet new lands await to be explored and charted, which include, but are not limited to, the development of linguistic data for mobile devices, the visualization of linguistic data, the creation of user-specific language data (for linguists, language teachers and learners and amateur users), etc.

Computer tools, powered by large quantities of linguistic data and advanced machine-learning algorithms, have provided a wide range of language services. Linggle, for instance, a web-scale linguistic search engine for words in context designed by Professor Jason S. Chang at the National Tsing Hua University (Taiwan), extracts collocations from large corpora to assist learners of English as a second language in writing (Boisson et al. 2013). Such services will be more common and more user-friendly in the future. Further, linguistic theories have a role to play when language resources are used, which will greatly benefit the resulting applications. Without theoretical analyses, as Levin and Song (1997) put it in the abstract of their paper, "corpus evidence is difficult to understand and systemize".

What do language resources mean to theoretical linguistics? This question may be answered by making a distinction between (1) corpus-driven linguistics, which aims to build theories exclusively on corpus data with no annotation, completely free from pre-corpus theoretical premises, and (2) corpus-based linguistics, which uses corpus annotation with the aim of testing and improving theories (Gries 2012). Obviously, the former presents an ideal blueprint for the future of corpus linguistics, while the latter depicts what it is like today.

There might be a day when NLP becomes fully mature and AI sufficiently powerful so that language resources can turn into the magic lamp of Aladdin to help people with language-related tasks (such as foreign language learning and translation), telling them what it is and how to do it. As Chomsky has argued, however, the point of science is to understand as much as one can about the nature of things to answer the question "Why?"(Andor 2004). Linguistic data and related computer tools nowadays, while helping to sketch what language is and how it works, provide little insight into the question of why language is the way it is. In this sense, the future of the science of language lies in a combination of human insights and language resources. It is time to promote the connection between introspection-based theoretical analysis and data-driven statistical analysis, rather than arguing against the value and potential of one another.

Notes

1 For example, the page of links (http://web.csulb.edu/~txie/pcr.htm) maintained by Dr Tianwei Xie, California State University, Long Beach, to help students who use the textbook *Practical Chinese Reader*. The collection includes 21 types of resources, 216 items in total.

2 More information can be found at https://en.wikipedia.org/wiki/Google_Ngram_Viewer; online search at https://books.google.com/ngrams; data available for download at http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

3 The research contributes to machine translation as well. Google Translate today, powered by an artificial neural network to increase fluency and accuracy, is still rendering the sentence "They love to read and be read to" as "他们喜欢阅读和阅读 *Tāmén xǐhuān yuèdú hé yuèdú*" "They love to read and read", showing much room for improvement in this regard.

4 For instance, verb-complement constructions like 说回 (那道栗子汤) *shuō huí* (*nà dào lìzǐtāng*) 'return to (the previous topic of the chestnut soup)' are found in the corpus, the complement being the directional verb 回 *huí* 'return' and the noun phrase 栗子汤 *lìzǐtāng* 'the chestnut soup' following the construction as the object. This is, however, rarely found in standard Chinese.

Further reading

Chomsky, Noam. 2000. *New horizons in the study of language and mind*. Cambridge: Cambridge University Press.

Fang, Alex C. Y., Yanjiao Li, Jing Cao, and Harry Bunt. 2019. Chinese multimodal resources for dialogue act analysis. In *The Routledge handbook of Chinese applied linguistics*, eds. Chu-Ren Huang, Zhuo Jing-Schmidt, and Barbara Meisterernst. London: Routledge.

Huang, Chu-Ren, and Nianwen Xue. 2019. Digital language resources and NLP tools. In *The Routledge handbook of Chinese applied linguistics*, eds. Chu-Ren Huang, Zhuo Jing-Schmidt, and Barbara Meisterernst. London: Routledge.
Lu, Qin. 2019. Computer and Chinese writing system. In *The Routledge handbook of Chinese applied linguistics*, eds. Chu-Ren Huang, Zhuo Jing-Schmidt, and Barbara Meisterernst. London: Routledge.
McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Abingdon: Routledge.
Newmeyer, Frederick J. 2003. Grammar is grammar and usage is usage. *Language* 79: 682–707.
Norvig, Peter. 2011. *On Chomsky and the two cultures of statistical learning*. Available at http://norvig.com/chomsky.html. Accessed 17 January 2018.
Sampson, Geoffrey. 2001. *Empirical linguistics*. London: Continuum.
Zhang, Jingwei, and Daming Xu. 2019. The impact of information and communication technology on Chinese language life. In *The Routledge handbook of Chinese applied linguistics*, eds. Chu-Ren Huang, Zhuo Jing-Schmidt, and Barbara Meisterernst. London: Routledge.

References

Andor, Jozsef. 2004. The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics* 1(1): 93–111.
Bai, Xiaojing, and Weidong Zhan 柏晓静， 詹卫东. 2006. Constraints on "bei" (passive) sentence in Chinese for machine translation of English passive sentence 汉语 "被" 字句的约束条件与机器翻译中英语被动句的处理. In *New advances in studies on passive sentences in Chinese 汉语被动表述问题研究新拓展*, ed. Fuyi Xing 邢福义, 1–17. Huazhong: Huazhong Normal University Press.
Boisson, Joanne, Ting-Hui Kao, Jian-Cheng Wu, Tzu-His Yen, and Jason S. Chang. 2013. Linggle: A web-scale linguistic search engine for words in context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 139–144. Stroudsburg, PA: Association for Computational Linguistics.
Feng, Shengli 冯胜利. 2015. Tone, intonation and sentence final particles in Chinese 声调、语调与汉语的句末语气词. *Essays on Linguistics*. *语言学论丛* 51: 52–79.
Feng, Zhiwei 冯志伟. 2006. *An introduction to the book* "*Corpora in Applied Linguistics*" 《*应用语言学中的语料库*》*导读*. Beijing: Beijing World Publishing Corporation.
Gries, Stefan Th. 2012. Corpus linguistics, theoretical linguistics, and cognitive psycholinguistics: Towards more and more fruitful exchanges. In *Corpus linguistics and variation in English: Theory and description*, eds. Joybrato Mukherjee and Magnus Huber, 41–63. Amsterdam: Rodopi.
Hong, Jia-Fei. 2014. Chinese near-synonym study based on the Chinese Gigaword corpus and the Chinese Learner Corpus. In *Chinese Lexical Semantics, 15th Workshop, CLSW 2014*, 329–340. Macao, China: Vol. 8922 of the series Lecture Notes in Computer Science, Springer International.
Hong, Jia-Fei, and Chu-Ren Huang. 2006. *Using Chinese Gigaword corpus and Chinese Word Sketch in linguistic research*. Paper presented at the Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, Wuhan, China.
Huang, Chu-Ren, and Nianwen Xue. 2019. Digital language resources and NLP tools. In *The Routledge Handbook of Chinese Applied Linguistics*, eds. Chu-Ren Huang, Zhuo Jing-Schmidt, and Barbara Meisterernst. London: Routledge.
Huang, Chu-Ren, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese Sketch Engine and the extraction of grammatical collocations. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 48–55. Stroudsburg, PA: Association for Computational Linguistics.
Institute of Linguistics, Chinese Academy of Social Science (CASS)中国社会科学院语言研究所. 2011. *Modern Chinese dictionary* (5th ed.) *现代汉语词典*（第5版). Beijing: Commercial Press.
Jing, Yingqi, and Haitao Liu. 2015. Mean hierarchical distance augmenting mean dependency distance. In *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015), 161–170. Uppsala, Sweden.
Levin, Beth, and Grace Song. 1997. Making sense of corpus data: A case study of verbs of sound. *International Journal of Corpus Linguistics* 2(1): 23–64.
Li, Yuming 李宇明. 2007. 关于 《中国语言学生活绿皮书》 [On green book on language situation in China]. *Applied Linguistics 语言文字应用* 1: 12–19.

Liu, Haitao, Richard Hudson, and Zhiwei Feng. 2009. Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory* 5(2): 161–175.

Ma, Qian 马千. 2011. *The investigation of definite expressions of Chinese "zhe (this)" and "na (that)" by comparing Chinese with English* 从英汉对比看汉语 "这"、"那"的定指表达. Master's thesis 硕士学位论文, Beijing: Library of Peking University.

Tao, Hongyin. 2000. Adverbs of absolute time and assertiveness in vernacular Chinese: A corpus-based study. *Journal of the Chinese Language Teachers Association* 35(2): 53–74.

Ye, Shumian 叶述冕. 2016. Typological correlations between tones, intonation and particles: A case study of polar questions 声调、语调、语气词之类型学相关性. *Essays on Linguistics 语言学论丛* 53: 336–363.

Yuan, Yulin, and Qiang Li 袁毓林，李强. 2014. How to use qualia structure to solve "tennis problem"? 怎样用物性结构知识解决"网球问题"? *Journal of Chinese Information Processing 中文信息学报* 28(5): 1–12.

Zhan, Weidong 詹卫东. 2013a. Chinese linguistics in the era of big data 大数据时代的汉语语言学研究. *Journal of Shanxi University (Philosophy & Social Science) 山西大学学报（哲学社会科学版）* 36(5): 70–77.

Zhan, Weidong 詹卫东. 2013b. Grammatical knowledge extraction from a large-scale Chinese treebank 基于大规模中文树库的汉语句法知识获取研究. In *Human language resources and linguistic typology – Papers from the 4th International Conference on Sinology 语言资讯与语言类型（中央研究院第四届国际汉学会议论文集）*, ed. Chiu-yu Tseng 郑秋豫, 239–267. Taipei: Institute of Linguistics, Academia Sinica.

Zhan, Weidong, Baobao Chang, and Shiwen Yu 詹卫东，常宝宝，俞士汶. 1999. Analysis on types of phrase boundary ambiguity in contemporary Chinese 汉语短语结构定界歧义类型分析及分布统计. *Journal of Chinese Information Processing 中文信息学报* 3: 9–17.

Zhan, Weidong, and Hongyin Tao 詹卫东，陶红印. 2016. A corpus approach to North American Chinese based on written media texts 北美书面汉语语法特点探析 – – 基于互联网中文文本的考察. *Global Chinese 全球华语* 2(1): 51–72.

Appendix

Resources		*Websites*	*Developers and maintainers*
国家语委现代汉语平衡语料库	Balanced Corpus of Modern Chinese	www.cncorpus.org	国家语言文字工作委员会 State Language Commission
国家语委古籍语料库	Balanced Corpus of Ancient Chinese	www.cncorpus.org	国家语言文字工作委员会 State Language Commission
北京语言大学汉语语料库	BCC Corpus online	http://bcc.blcu.edu.cn	北京语言大学 Beijing Language & Culture University
北京语言大学HSK动态作文语料库	HSK Learner Corpus of Composition Texts	http://bcc.blcu.edu.cn/hsk	北京语言大学 Beijing Language & Culture University
哈工大中文篇章关系语料	Chinese Discourse Annotated Corpus of Harbin Institute of Technology	http://ir.hit.edu.cn/hit-cdtb/	哈尔滨工业大学 Harbin Institute of Technology
哈工大语义依存树库	Chinese Dependency Treebank of Harbin Institute of Technology	www.ltp-cloud.com/intro/#sdp_how	哈尔滨工业大学 Harbin Institute of Technology
清华大学汉语句法树库	Chinese Syntactic Treebank of Tsinghua University	http://cslt.riit.tsinghua.edu.cn/~qzhou/papers/TCTScheme.pdf	清华大学 Tsinghua University
情感词汇本体库	Emotion Ontology of Chinese Words	http://ir.dlut.edu.cn/EmotionOntology Download	大连理工大学 Dalian University of Technology
北京大学现代汉语语法信息词典	*The Grammatical Knowledgebase of Contemporary Chinese*	http://icl.pku.edu.cn/icl_groups/syntac-dictn.asp	北京大学 Peking University
北京大学中国语言学研究中心语料库	CCL Corpus online	http://ccl.pku.edu.cn:8080/ccl_corpus	北京大学 Peking University
知网	HowNet	www.keenage.com/html/e_index.html	中国中文信息学会（董振东） Chinese Information Processing Society of China (Prof Zhendong Dong)
中文语言资源联盟	Chinese LDC	www.chineseldc.org/	中国中文信息学会 Chinese Information Processing Society of China

Resources		*Websites*	*Developers and maintainers*
国家语言资源监测与研究中心平面媒体分中心	National Language Resource Monitoring & Research Center	http://dcc.blcu.edu.cn/main.action	北京语言大学 Beijing Language & Culture University
国家语言资源监测与研究中心有声媒体分中心	National Language Resource Monitoring & Research Center	http://ling.cuc.edu.cn/RawPub/	中国传媒大学 Communication University of China
国家语言资源监测与研究中心网络媒体分中心	National Language Resource Monitoring & Research Center	http://nlp.ccnu.edu.cn/	华中师范大学 Central China Normal University
国家语言资源监测与研究中心教育教材语言分中心	National Language Resource Monitoring & Research Center	http://ncl.xmu.edu.cn/	厦门大学 Xiamen University
国家语言资源监测与研究中心海外华语研究分中心	National Language Resource Monitoring & Research Center	http://huayu.jnu.edu.cn/	暨南大学 Jinan University
国家语言资源监测与研究中心少数民族语言分中心	National Language Resource Monitoring & Research Center of Minority Languages	http://nmlr.muc.edu.cn/	中央民族大学 Minzu University of China
早期粤语口语文献数据库	Early Cantonese Colloquial Texts: A Database	http://pvs0001.ust.hk/Candbase/	香港科技大学 Hong Kong University of Science and Technology
早期粤语标注语料库	Early Cantonese Tagged Database	http://pvs0001.ust.hk/WTagging/	香港科技大学 Hong Kong University of Science and Technology
香港律政司双语法例资料系统	BLIS (Bilingual Laws Information System)	http://translate.legislation.gov.hk/gb/www.legislation.gov.hk/blis/chi/index.html	香港律政司 Department of Justice, The Government of Hong Kong SAR
香港城大泛华语共时同题语料库	LIVAC (Linguistic Variation in Chinese Speech Communities)	www.livac.org	香港城市大学 City University of Hong Kong
香港双语儿童语言资料库	The Hong Kong Bilingual Child Language Corpus	www.cuhk.edu.hk/lin/home/bilingual.htm	香港中文大学 Chinese University of Hong Kong
汉英平行语料库	English-Chinese Parallel Concordance	http://ec-concord.ied.edu.hk/paraconc	香港教育大学 Education University of Hong Kong

Resources		*Websites*	*Developers and maintainers*
数位学习国家型科技研究计划（2004－05）语料库集	Taiwan e-Learning and Digital Archives Programme	http://elearning.ling.sinica.edu.tw/resources.html	台湾中央研究院语言学研究所 Institute of Linguistics, Academia Sinica
语言典藏计划（第一期、第二期）语料库集	Taiwan e-Learning and Digital Archives Programme (Phase I, II)	http://languagearchives.sinica.edu.tw/cht/index.php.html	台湾中央研究院语言学研究所 Institute of Linguistics, Academia Sinica
中文词汇网络	Chinese WordNet	http://cwn.ling.sinica.edu.tw/	台湾中央研究院语言学研究所 Institute of Linguistics, Academia Sinica
中英双语本体知识库	The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW)	http://bow.ling.sinica.edu.tw/intro/bow_ebg_cont.html	台湾中央研究院语言学研究所、资讯科学研究所 Institute of Information Science, Institute of Linguistics, Academia Sinica
中文词知识库	Chinese Lexicons	http://ckip.iis.sinica.edu.tw/CKIP/index.htm	台湾中央研究院资讯科学研究所 Institute of Information Science, Academia Sinica
汉字知识本体	Hantology	http://hantology.ling.sinica.edu.tw/index.htm	台湾大学 Taiwan University
搜文解字		http://words.sinica.edu.tw	台湾国科会 National Science Council
	Korean-Chinese Parallel Corpus	http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus	Korea Advanced Institute of Science and Technology, KAIST
	CoreNet	http://semanticweb.kaist.ac.kr/home/index.php/CoreNet	Korea Advanced Institute of Science and Technology, KAIST
	ASPEC-JC (Asian Scientific Paper Excerpt Corpus-JC)	http://lotus.kuee.kyoto-u.ac.jp/ASPEC/	Japan Science and Technology Agency (JST), National Institute of Information and Communications Technology (NICT)

Resources		*Websites*	*Developers and maintainers*
	NICT Japanese-Chinese Parallel Corpus	http://universal.elra.info/product_info.php?cPath=42_43&products_id=2044	National Institute of Information and Communications Technology (NICT), Japan
新加坡 多语语料库	NTU Multilingual Corpus	http://compling.hss.ntu.edu.sg/ntumc/	Nanyang Technological University, Singapore
汉语开放词网	Chinese Open Wordnet	http://compling.hss.ntu.edu.sg/cow/	Nanyang Technological University, Singapore
语言资源联盟（LDC）中文部分	Linguistic Data Consortium LDC	www.ldc.upenn.edu/	University of Pennsylvania, USA
UCLA 汉语书面语语料库	The UCLA Written Chinese Corpus	www.lancaster.ac.uk/fass/projects/corpus/UCLA/default.htm	University of California Los Angeles, UCREL of Lancaster University, USA
洛杉矶中文学习中心语音学习资料	Learn Chinese Online via Podcast & MP3	http://chinese-school.netfirms.com/learn-Chinese-online.html	Los Angeles Chinese Learning Center, USA
	The GDELT Project	www.gdeltproject.org/	Google GDELT Project, USA
汉语综合语料库	A collection of Chinese corpora and frequency lists	http://corpus.leeds.ac.uk/query-zh.html	Leeds University, UK
汉英平行语料库	The Babel English-Chinese Parallel Corpus	www.lancaster.ac.uk/fass/projects/corpus/babel/babel.htm	Lancaster University, UK
兰开斯特大学中文语料库	The Lancaster Corpus of Mandarin Chinese	www.lancaster.ac.uk/fass/projects/corpus/LCMC	Lancaster University, UK
兰开斯特洛杉矶汉语口语语料库	The Lancaster Los Angeles Spoken Chinese Corpus (LLSCC)	www.lancaster.ac.uk/fass/projects/corpus/LLSCC/	Lancaster University, UK
PDC2000(2000年《人民日报》全年）语料库	The PDC2000 Corpus of Chinese News Text Word Sketch Engine	www.lancaster.ac.uk/fass/projects/corpus/pdc2000/ www.sketchengine.co.uk/	Lancaster University, UK Lexical Computing, UK