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

Introduction

Bianca Basciano
Università Ca' Foscari Venezia, Italia

Franco Gatti
Università Ca' Foscari Venezia, Italia

Anna Morbiato
Università Ca' Foscari Venezia, Italia; The University of Sydney, Australia

In the past decades, corpus-based research has been gaining momentum in contemporary linguistics. While corpora, intended as large collections of naturally occurring texts, have always existed, rapid advances in computation and technology have provided tools for faster and more effective corpus construction and consultation. Chinese makes no exception: corpus data are now considered among the main resource for many linguists, while large-scale surveys are beginning to be taken as an important tool for linguistic investigation.

Among the reasons beyond the increasing number of corpus-based studies is the availability of "a myriad of large and publicly available Chinese corpora" (Xu 2015, 219), which include general purpose corpora, such as the CCL (Centre for Chinese Linguistics, Peking University) corpus or the BCC (Beijing Languages and Cultures University) corpus, interlanguage corpora, such as the BLCU International Corpus of Learner Chinese, and specialised corpora, such as the ZHTenTen simplified Chinese corpus mounted at Sketch Engine, the LDC (Linguistic Data Consortium at UPenn) or the ELRA (European Language Resources Association). Smaller, genre- or domain-specific corpora, such as e.g. the Leiden Weibo Corpus or the *Renmin Ribao* 'People's Daily' database, are also growing in number. Other resources include multilingual corpora and databases – e.g. a number

of English-Chinese parallel corpora, translational Chinese corpora, Chinese dialects databases and corpora of ethnic languages in China.

The great advantage of corpora lies in the fact that they offer access to large amounts of authentic, naturally occurring linguistic data produced by a variety of speakers or writers, thus providing more robust, statistically significant foundations for linguistic accounts and analyses. There is now considerable emphasis on the reliability of linguistic materials: several scholars stress the need for a shift to a more empirical mode of investigation, as rigorous theoretical advances need to be "grounded in solid empirical data" (Jing-Schmidt 2013, 1). A further advantage is that corpus queries may also reveal the statistical relevance of a specific linguistic phenomenon, e.g. a lexical item or a grammatical pattern, as well as possible changes or developments of its behaviours over time. Moreover, corpus queries may also allow searching for significant interactions between domain variables (Wallis, Nelson 2001). Finally, these tools may help reveal new words or patterns that were previously unobservable or, else, regarded as non-existent or marginal. In short, corpora allow qualitative and quantitative, synchronic and diachronic investigations of the language, providing factual, frequency, and interaction evidence for linguistic analyses (Wallis 2019). They not only offer new insights within the core subfields of linguistics – including syntax, semantics and lexicography, pragmatics and language use, information structure – but also provide precious material for disciplines such as language acquisition, with the analysis of learners' corpora and interlanguage development, or sociolinguistics, with synchronic and diachronic studies on language and society, socio-linguistic comparison, as well as the development of buzzwords in social media and the Internet.

The past decade has seen the rapid development of corpus-based research in many aspects of Chinese language and linguistics. One of the most popular types of research is the compilation of frequency character/word lists (Xu 2015): after Li Jinxi's *A Statistical Analysis of Basic Chinese Vocabulary* (1922), lexical studies received increasing interest, with many scholars applying corpus tools to all aspects of lexicography, including selecting words to be included in a dictionary on a statistical basis, identifying word senses, ordering of polysemous and homograph items, as well as determining word classes and singling out illustrative examples of words' uses (see McEnery, Xiao 2016, 442). Among the most recent lexical frequency and word list projects, there are the latest national Chinese character list, i.e. the 通用规范汉字表 *Tōngyòng Guīfàn Hànzì Biǎo* (A General Service List of Chinese Characters), released in 2013, and Xiao, Rayson and McEnery's (2009) *A Frequency Dictionary of Mandarin Chinese* (see McEnery, Xiao 2016 for a review). Corpus-based researches on second language acquisition and interlanguage development have also been increasing over the last couple of decades, with early projects at

BLCU now developed into the BLCU International Corpus of Learner Chinese, followed by other studies (Tao 2008, 2009; Xiao 2007; Zou, Smith, Hoey 2016, *inter alia*; for a review, see Xu 2015; McEnery, Xiao 2016; Zhang, Tao 2018). On the other hand, scholars agree that corpus-based sentential/grammatical level research is practically negligible if compared with lexical studies, although it is now receiving increasing attention with the introduction of more sophisticated query tools. For example, there have been some innovative corpus studies on morphological aspects of Chinese, e.g. on compounds and affixes (Sproat, Shih 1996; Nishimoto 2003; Arcodia, Basciano 2012) and on 离合词 *líhécí* 'separable words' (Siewierska, Xu, Xiao 2010; Wang C. 2001, Wang H. 2011). With respect to syntax, remarkable insights have been gained by scholars using corpora on syntactic patterns and behaviours of, e.g. adjectives (Thompson, Tao 2010), adverbial clauses (Wang 2006), and verbal coercion (Tao 2000). Interesting work has also been done on discourse/pragmatics (Jing-Schmidt, Kapatsinsky 2012). Contrastive studies also constitute a promising line of research, with main works done on the differences between English and Chinese (Xiao, McEnery 2008, 2010). Other significant areas of inquiry include corpus and database construction (Zhan 2019) and historical linguistics (Halliday 1959; Cook 2011; Ji 2010); for an overview, see Xu (2015). However, apart from these notable exceptions, Chinese corpus-based theoretical linguistics studies are scarce and by no means the mainstream (Xu 2015), partly due to the technological and methodological limitations connected with corpus interrogation. McEnery and Xiao (2016) also hold that research in corpus-based descriptive grammar in Chinese is rather sporadic and fragmentary, and has focused on specific linguistic features of interest to individual researchers.

This volume wants to contribute to filling this gap and stems from the idea that a lot can still be done: issues that have not received a commonly accepted account may benefit from corpus-based investigation conducted from a different angle, qualitative and/or quantitative; second, corpora may reveal linguistic phenomena, patterns and constructions that have not yet been investigated, thus enriching our knowledge of grammar; finally, new corpora or corpus-tagging methods that allow more precise analyses in specific research fields, ranging from diachronic linguistics to sociolinguistics, syntax and pragmatics, can be identified and suggested for future lines of research.

Studies presented in this volume are both quantitative and qualitative, as well as synchronic and diachronic, and are grounded in the tenet that corpora provide a more robust, statistically significant foundation for linguistic analyses. As corpus linguistics is not a monolithic, consensually agreed set of methods and procedures (McEnery, Hardie 2011), differences inevitably exist regarding approaches and methodologies in the different contributions, which may be both

discipline-specific and also due to the different aim and focus of each study. The contributions provide different insights not only into the potential of using corpora as tools allowing access to authentic language material, but also into the challenges involved in corpus interrogation, analysis, and building. All in all, they contribute to answering three fundamental questions: how can corpora improve current theoretical accounts of Chinese grammar in general? What do corpora reveal about the statistical relevance of linguistic phenomena and constructions? What are the limitations and the drawbacks of using corpora to investigate Chinese languages?

As reflected in the five sections of the volume, the contributions cover different fields of linguistics, including syntax and pragmatics, semantics, morphology and the lexicon, sociolinguistics, and corpus building.

The first section explores issues in Chinese syntax and pragmatics. Tao, Jin and Zhang's paper proposes an investigation of manner and state complement constructions combining corpus-based and corpus-driven methods, based on a corpus of written Chinese, offering both a theoretical account and an exploration of the implications for Chinese L2 learning. The study highlights preferred forms and functions of Manner/State Complement Constructions: monosyllabic verbs, basic action verbs, or psychological state verbs tend to co-occur with complements of adjectival, clausal, or idiomatic expressions. The authors conclude that Manner/State Complement Constructions are an assessment device indexing speaker evaluative stance, and that the loaded affective meanings account for the larger and more complex forms than their standard counterparts.

Morbiato provides quantitative and qualitative evidence of the existence of indefinite NPs in the sentence-initial and preverbal position, thus ruling out strict associations between definiteness, givenness, and the sentence-initial position and related restrictions often referred to in the literature. She examines big-size, generalised corpora, such as the PKU CCL corpus (Peking University), the BCC corpus (Beijing Language and Culture University), and the ZHTenTen (Stanford Tagger) corpus mounted at Sketch Engine. Her statistical data show that this phenomenon is neither rare nor marginal. Furthermore, they reveal that animate indefinites are significantly more likely to occur sentence-initially, while locatability and partitivity are frequent traits of inanimate SIIs. Finally, it singles out and discusses a new pattern featuring a proper noun introduced by the indefinite marker '一 *yī* + CLASSIFIER', thus confirming that corpora indeed contribute towards a more complete understanding of a language system by allowing to single out new, previously underdescribed linguistic patterns and phenomena.

Tantucci and Wang explore the V-过 *guo* construction by examining its evidential *versus* experiential usages in two comparable writ-

ten corpora, i.e. the Lancaster Corpus of Mandarin Chinese and the UCLA corpus of written Mandarin. The results of this study shed light on the relationship between the formal and functional categories of the V-过 *guo* construction and the textual environment in which it occurs, showing that specific genres and textual environments favour the evidential usage of 过 *guo* and that evidentiality is an important grammatical category of documentary, factual and academic prose. This study also shows that the categorial separation between evidential and experiential usages of the construction is a result of features underpinning form, usage and 'contextual situatedness'. The authors conclude that evidentiality emerges from specific intersections among these three dimensions and from distinctive illocutional concurrences of conventionalized behaviour.

The second section is devoted to semantic studies. Shi, Liu and Jing-Schmidt present a usage-based, quantitative and qualitative corpus investigation of action metaphors involving manual object manipulation. Two transitive constructions, [抓紧 *zhuājǐn* 'grab tightly, clutch' NP] and [把住 *bǎzhù* 'grasp firmly' NP], and a causative construction, [把 *bǎ* NP 捧 *pěng* COMPL] 'lift NP with deliberation' (with a metaphoric sense), are examined: results reveal that the former systematically imply a keen sense of urgency and/or importance, while the latter involves over-promotion of an undeserving entity. The study highlights the methodological importance of quantitative studies in establishing the conventionality, productivity, and semantic subclassification of metaphors encoded in syntactic patterns. It has both implications for theoretical hypotheses regarding the embodiment of conceptualisation and for language learning and teaching.

The contribution by Sparvoli focuses on modality, in particular on the factuality reading triggered by Chinese modals in past contexts. Through a corpus-based investigation, conducted in the English Chinese Parallel Concordancer, published by the Hong Kong Institute of Education, the author tests the hypothesis that deontic modals trigger counterfactual inference, while anankastic/goal-oriented modals either trigger an actuality entailment effect or a generic non-factual reading. The results of her investigation confirm the crucial role played by the deontic vs. anankastic contrast in the marking of factuality in Chinese, showing a gradient cline, from anankastic/goal-oriented modals to deontic modals, along which the factuality value decreases. The two extreme poles of the cline get a unique reading, i.e. past counterfactual for pure deontic modals and factual for strong anankastic modals. Finally, some pedagogical implications are discussed.

Boaretto and Castello propose a corpus-based study of Chinese modality by comparing the English and Chinese versions of Pope Francis' second encyclical *Laudato Si'*, focusing on different areas of modality, i.e. prediction/volition/intention, lack of possibility/abil-

ity/permission, and obligation. Meaningful translation correspondences are investigated to define their semantic space and detect possible cases of explicitation. While corpus data confirm predictable parallel expressions such as *will* and 会 *huì*, *cannot* and 不能 *bù néng*, they also reveal new correspondences, such as no overt modal expression in English and 会 *huì*, or *cannot* and 无法 *wúfǎ*. Overall, the study highlights how the translation of highly grammaticalised items undergoes a process of interpretation and adaptation: some translation choices are due to the translator's attempt to make the text explicit and to adapt it to the target culture. The corpus-based approach adopted reveals a network of semantically connected modal expressions and helps to identify the linguistic choices made by the writer and the translator to convey the intended semantic meanings. The authors point out that, while parallel concordancing software could help speed up this type of analysis, human scrutiny and judgement are still needed.

The third section proposes research into the lexicon and morphology of Chinese. Specifically, Dosedlová and Lu propose a corpus-based study on near-synonymy of classifiers: in Chinese there are many classifiers which are near-synonymous and interchangeable in some contexts. In particular, the study investigates two near-synonymous classifiers, i.e. 棵 *kē* and 株 *zhū*, based on co-varying collexeme analysis, which belongs to collostrucional methods (i.e. corpus-based quantitative methods which measure mutual attraction between lexemes and constructions), and on Euclidean distance. Such an approach allows to obtain a clearer picture on the co-occurrence of certain classifiers with certain nouns and on different usages. However, the authors suggest that it is highly recommendable to combine different methodological approaches for the analysis of near synonymy, in order to obtain a more comprehensive picture, able to reveal different aspects of the phenomenon.

The contribution by Basciano and Bareato focuses on word-formation, specifically, on new word-formation patterns emerged in the last few decades under the influence of foreign languages and netspeak. The authors present a corpus-based investigation on three emerging suffixes, i.e. 族 *zú*, 党 *dǎng*, and 客 *kè*, all forming nouns indicating persons with certain characteristics or behaviour, or doing a certain activity, examining neologisms drawn from the following sources: the 新世纪新词语大词典 *Xin shiji xinciyu da cidian* (New Century Comprehensive Dictionary of Neologisms), the Leiden Weibo Corpus, and the Buzzwords section of the *Shanghai Daily*. After describing the three word-formation patterns, the paper describes their evolution over time, and the semantic shift and meaning generalisation characterising their grammaticalization path. The study also proposes an analysis of productivity measures for the three word-formation patterns and discusses their diffusion in Chinese.

The fourth section explores applications of corpus tools to the investigation of sociolinguistic aspects. Specifically, Chin proposes a novel use of the Corpus of Mid-20th Century Hong Kong Cantonese, i.e. as a window on Hong Kong society, and specifically its family structure and marital life. It consists of a corpus-based sociolinguistic investigation of kinship terms and terms related to marriage, which reveals significant differences in family structure as compared to contemporary Hong Kong society.

The fifth section tackles issues on corpus and database construction. Zhan et al. present their work in progress and the challenges encountered in the creation of a Chinese constructicon provisionally named CCL-CxnBank. The project has been carried out since 2015 by the Center for Chinese Linguistics of Peking University and, at the moment, the constructicon includes more than 1,000 constructions and records their syntactic, semantic, and pragmatic information, as well as synonymy, antonymy, and hyponymy/hypernymy relations. In addition, the project includes the annotation of a corpus collecting instances of various usages of the constructions in real contexts: the corpus annotates the internal structure and the subjective attitude meaning of each construct, in order to provide a comprehensive description of the actual usages of the constructions.

Lastly, Anderl presents some reflections on the Database of Medieval Chinese Texts, an international and collaborative project, drawing on the expertise of specialists in various fields, the main partners being Ghent University and Dharma Drum Institute of Liberal Arts (Taiwan). The database collects manuscript texts, with a focus on the period between ca. 700 and 1000 CE. While there is a variety of digital databases for premodern Chinese texts, specialised databases on non-canonical manuscripts are still very rare and provide rather limited information. Therefore, this project is very valuable, since it aims at providing high-quality digital editions of Late Medieval Chinese key texts, which are of great importance for research on early colloquial grammatical markers and syntactic constructions, also developing an analytical apparatus. The paper presents the technical framework, the reference data collections, the process of digitalisation of the texts, the various modules of the database, and proposes some reflections. The paper also discusses the importance of the database as a pedagogical tool.

We would like to thank all the anonymous reviewers for their precious help. We would also like to express our heartfelt gratitude to Magda Abbiati and Federico Greselin for their generous support. Lastly, we wish to express our gratitude to the editorial staff of Edizioni Ca' Foscari.

Bibliography

Arcodia, G. F.; Basciano, B. (2012). "On the Productivity of the Chinese Affixes –兒 *–r*, –化 *–huà* and –頭 *–tou*". *Taiwan Journal of Linguistics*, 10, 89-118. http://dx.doi.org/10.6519/TJL.2012.10(2).3.

Cook, A. (2011). "Recent Developments in the Use of the Plural Marker *Men* in Modern Standard Chinese in Taiwan". *Chinese Language and Discourse*, 2(1), 80-98. https://doi.org/10.1075/cld.2.1.04coo.

Ji, M. (2010). "A Corpus-Based Study of Lexical Periodization in Historical Chinese". *Literary and Linguistic Computing*, 25(2), 199-213. https://doi.org/10.1093/llc/fqq002.

Jing-Schmidt, Z. (2013). *Increased Empiricism: Recent Advances in Chinese Linguistics*. Amsterdam: John Benjamins.

Jing-Schmidt, Z.; Kapatsinsky, V. (2012). "The Apprehensive: Fear as Endophoric Evidence and Its Pragmatics in English, Mandarin, and Russian". *Journal of Pragmatics*, 44, 346-73. https://doi.org/10.1016/j.pragma.2012.01.009.

Halliday, M. (1959). *The Language of the Chinese "Secret History of the Mongols"*. Oxford: Basil Blackwell.

Li J. 黎锦熙 (1922). "Guoyu zhong jiben yuci de tongji yanjiu" 国语中基本语词的统计研究 (Statistical Considerations of Basic Vocabulary in Chinese). *Guowen xuehui congkan*, 1(1), 81-4.

McEnery, T.; Hardie, A. (2011). "What Is Corpus Linguistics?". McEnery, T.; Hardie, A. (eds), *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 1-24.

McEnery, T.; Xiao, R. (2016). "Corpus-Based Study of Chinese". Chan, S. (ed.), *The Routledge Encyclopedia of the Chinese Language*. New York: Routledge, 438-51.

Nishimoto, E. (2003). "Measuring and Comparing the Productivity of Mandarin Chinese Suffixes". *Computational Linguistics and Chinese Language Processing*, 8(1), 49-76.

Siewierska, A; Xu, J.; Xiao, R. (2010). "Bang-le Yi Ge Da Mang (Offered a Big Helping Hand): A Corpus Study of the Splittable Compounds in Spoken and Witten Chinese". *Language Sciences*, 32(4), 464-87. https://doi.org/10.1016/j.langsci.2009.08.002.

Sproat, R.; Shih, C. (1996). "A Corpus-Based Analysis of Mandarin Nominal Root Compounds". *Journal of East Asian Linguistics*, 5, 49-71. https://doi.org/10.1007/BF00129805.

Tao H. 陶红印 (2000). "Cong 'Chi' Kan Dongci Lunyuan Jiegou de Dongtai Tezheng" 从"吃"看动词论元结构的动态特征 ('Eating' and Emergent Argument Structure). *Yuyan Yanjiu*, 20(3), 21-38.

Tao, H. (2008). "The Role of Corpora in Chinese Language Teaching and Teacher Education". Duff, P.; Lester, P. (eds), *Issues in Chinese Language Education and Teacher Development*. Vancouver: Centre for Research in Chinese Language and Literacy Education, University of British Columbia, 90-102.

Tao, H. (2009). "Core Vocabulary in Spoken Mandarin and the Integration of Corpus- Based Findings into Language Pedagogy". Xiao Y. (ed.), *Proceedings of the 21st North American Conference on Chinese Linguistics*. Smithfield (Rhode Island): Bryant University, 13-27.

Thompson, S.; Tao, H. (2010). "Conversation, Grammar, and Fixedness: Adjectives in Mandarin Revisited". *Chinese Language and Discourse*, 1(1), 3-30.

Wallis, S. (2019). "Grammar and Corpus Methodology". Aarts, B.; Bowie, J.; Popova, G. (eds), *The Oxford Handbook of English Grammar*. Oxford; New York: Oxford University Press, 59-83.

Wallis, S.; Nelson, G. (2001). "Knowledge Discovery in Grammatically Analysed Corpora". *Data Mining and Knowledge Discovery*, 5(4), 305-35.

Wang C. 王春霞 (2001). "Jiyu yuliaoku de liheci yanjiu" 基于语料库的离合词研究 (A Corpus-Based Study of Splittable Compounds) [MA dissertation]. Beijing: Beijing Language and Culture University.

Wang H. 王海峰 (2011). *Xiandai Hanyu liheci lixi xingshi gongneng yanjiu* 现代汉语离合词离析形式功能研究 (A Functional Study of the Split Forms of Separable Words in Modern Chinese). Beijing: Peking University Press.

Wang, Y. (2006). "The Information Structure of Adverbial Clauses in Chinese Discourse". *Taiwan Journal of Linguistics*, 4(1), 49-88.

Xiao, R. (2007). "What Can SLA Learn From Contrastive Corpus Linguistics? The Case of Passive Constructions in Chinese Learner English". *Indonesian Journal of English Language Teaching*, 3(2), 1-19.

Xiao, R.; McEnery, T. (2008). "Negation in Chinese: A Corpus-Based Study". *Journal of Chinese Linguistics*, 36(2), 274-330. https://www.jstor.org/stable/23756111.

Xiao, R.; McEnery, T. (2010). *Corpus-based Contrastive Studies of English and Chinese*. London/New York: Routledge.

Xiao, R.; Rayson, P.; McEnery, T. (2009). *A Frequency Dictionary of Mandarin Chinese: Core Vocabulary for Learners*. London: Routledge.

Xu, J. (2015). "Corpus-Based Chinese Studies: A Historical Review From the 1920s to the Present". *Chinese Language and Discourse*, 6(2), 218-44. https://doi.org/10.1075/cld.6.2.06xu.

Zhan W. 詹卫东 (2019). "Beijing Daxue CCL Yuliaoku de Yanzhi" 北京大学CCL语料库的研制. *Yuliaoku yuyanxue*, 6(1), 71-86.

Zhang, J.; Tao H. (eds) (2018). *Corpus-Based Research in Chinese as a Second Language*. London: Routledge.

Zou, B.; Smith, S.; Hoey M. (eds) (2016). *Corpus linguistics in Chinese contexts*. New York: Palgrave Macmillan.