Computational and Corpus Approaches to Chinese Language Learning: An Introduction

Xiaofei Lu and Berlin Chen

Abstract In this introductory chapter, we first provide a discussion of the rationale and objectives of the book. We then offer a brief review of the body of corpus linguistics research that intersects with Chinese language pedagogy and acquisition. This is followed by an overview of the state of the art of research in computational linguistics and natural language processing that pertains to Chinese language teaching, learning, and assessment. We conclude with a description of the organization of the book.

1 Rationale and Objectives of the Book

The past a few decades have witnessed remarkable progress in the construction of large corpora of spoken and written language produced by both first language (L1) and second language (L2) speakers and writers. In addition to their proven value in a vast range of linguistic research (McEnery and Hardie 2011), language corpora have increasingly been used to inform second and foreign language teaching, learning, and assessment (Aijmer 2009). The plethora of language corpora have also facilitated the development of natural language processing (NLP) technologies with various language understanding and production capabilities (Jurafsky and Martin 2008). Two types of NLP technologies that are especially useful in second and foreign language pedagogy and research are computational tools designed to automate corpus annotation and analysis at specific linguistic levels (Lu 2014) and educational NLP applications designed to facilitate different aspects of language teaching and learning or to assist with the assessment of different aspects of language production (Lu 2018).

X. Lu (✉)
The Pennsylvania State University, University Park, USA
e-mail: xxl13@psu.edu

B. Chen
National Taiwan Normal University, Taipei, Taiwan
e-mail: berlin@ntnu.edu.tw

X. Lu and B. Chen (eds.), *Computational and Corpus Approaches to Chinese Language Learning*, Chinese Language Learning Sciences,
https://doi.org/10.1007/978-981-13-3570-9_1

Second and foreign language teachers, learners, and testers, particularly those who practice computer-assisted language learning (CALL), stand to benefit from computational and corpus resources, tools, and methods in numerous productive ways. For example, different types of first or expert language corpora can serve as resources for teaching material selection and adaptation (Jin and Lu 2018) and for deriving pedagogically relevant lists of words, collocations, and various kinds of multiword expressions (Xiao et al. 2009). Corpora also allow for the extraction of large quantities of authentic examples of language use that can be utilized by teachers to explicate patterns of language use or by learners to discover and generalize patterns of language use (Flowerdew 2009). Analyses of learner corpora can help teachers identify characteristics, patterns, as well as individual variation of learner language use, acquisition, and development, which in turn can be used to inform the adjustment and individual customization of the emphasis of pedagogical intervention (Díaz-Negrillo et al. 2013). Computational tools designed to automate corpus annotation and analysis can be used to dramatically increase the scale and efficiency of these and other ways to exploit corpora for pedagogical purposes. At the same time, intelligent computer-assisted language learning (ICALL) systems that integrate various types of NLP capabilities have the potential to enhance learning in a myriad of ways (Lu 2018). For example, ICALL systems may automatically analyze input texts at various linguistic levels to facilitate text selection and sequencing, textual enhancement, and visualization of relationships and differences among different linguistic concepts and elements. They may also integrate technologies to automatically diagnose errors, problems, and challenges in learners' spoken or written responses to learning tasks and provide learners with individualized feedback and assistance in real time. Finally, computational systems designed to automatically assess learners' spoken or written performance can be of considerable help to language testers and teachers alike, as they help alleviate the time- and labor-intensiveness involved in scoring learner task performance, particularly in large-scale and time-sensitive assessment scenarios.

While a large proportion of computational and corpus linguistics research to date has centered on English, much progress has been made in constructing Chinese corpora oriented towards Chinese language pedagogy as well as in developing computational tools for annotating and analyzing Chinese texts and educational NLP applications for Chinese language learning and assessment. The goal of this volume is to bring together some of the most recent theoretical, empirical, methodological, and technological developments in applying computational and corpus approaches to teaching, learning and assessing Chinese as a second or foreign language. Given that research efforts in this area are still emerging, the volume does not focus on a specific, targeted aspect of Chinese language teaching and learning, nor does it cover the full range of topics in this area. Nevertheless, it is our hope that the insights from the studies included here will help advance the state of the art in theorizing, designing, and implementing research and practice in corpus-informed and NLP-enabled Chinese language teaching, learning, and assessment.

In the rest of this chapter, we first provide a brief review of the body of corpus linguistics research that intersects with Chinese language pedagogy and acquisition. This is followed by an overview of the state of the art of research in computational linguistics that pertains to Chinese language teaching, learning, and assessment. We conclude with a description of the organization of the book.

2 Corpus Linguistics and Chinese Language Learning

Corpus resources and findings from corpus-based research into L1 and L2 use and acquisition have found increasingly broad applications in language teaching and learning. We do not intend to cover the full spectrum of such applications here; nor do we intend to repeat the same information that has been covered in the body of works on the same subject for English and other languages (e.g., Aijmer 2009; Flowerdew 2009). Instead, we focus on five specific types of applications in the context of Chinese as a second or foreign language that are touched upon in some of the chapters in this volume.

A long-standing use of corpus resources has been the derivation of pedagogically useful lists of words, multiword expressions, and the like (e.g., Xu this volume). Work in this area requires large-scale general-purpose or specialized corpora that are representative of the language being targeted (in terms of mode, register, genre, etc.) and that have been annotated with necessary linguistic information (e.g., parts of speech and lemmas of the words), a clear operationalization of the linguistic items or units to be extracted (e.g., collocations, phrasal verbs, lexical bundles, etc.), as well as a set of procedures and corpus statistics to be used to extract, filter, and organize the linguistic items or units to be included in the list. A good example of this line of research is Xiao et al.'s (2009) work on deriving a frequency dictionary of Mandarin Chinese from a large corpus of close to 50 million Chinese words. In this volume, Chen and Tao describe several efforts in deriving lists of academic vocabulary and lexical bundles from specialized corpora of academic Chinese.

Another well-established use of corpus resources has been in-depth analyses of specific language phenomena using corpus-based methods. Such analyses can generate valuable insights into the nature of the linguistic phenomena in question, which in turn can be used to inform the teaching and learning of those phenomena. Two chapters in this volume report work that exemplifies new advances in this area. Zhang shows how corpus-based analysis and visualization of differences among Chinese near-synonyms using Correspondence Analysis can yield highly useful information for teaching and learning Chinese near-synonyms. Gong and Hong illustrate the usefulness of corpus-based analysis of neologisms originated from Chinese new media in helping Chinese learners acquire not only lexical and morphological knowledge but also sociocultural knowledge that underlies the construction and emergence of such neologisms.

Parallel corpora, i.e., corpora consisting of texts in the original language as well as their translations in one or more other languages, have found their way into corpus-

informed language pedagogy as well. As detailed by Bluemel in this volume, while such corpora were initially designed to facilitate research in contrastive linguistics and translation studies, they have been shown to be particularly useful in helping learners develop conceptual knowledge of complex linguistic constructions in Chinese. His novel design of the Parallel Corpus Teaching Tool and successful application of it in beginning-level Chinese classes break new ground in theory-informed integration of parallel corpora in Chinese language pedagogy.

Text corpora and various types of corpus-derived lists are now also being used as benchmarks to inform reading text selection and adaptation. This line of research involves establishing benchmarks of text complexity features of collections of texts of different grade, readability, or difficulty levels and using such benchmarks to assist the processes of selecting and/or adapting additional reading texts for learners at different proficiency levels (Jin and Lu 2018). In this volume, Bo et al. describe the development of an online text adaptation system that uses linguistic features mined from a large corpus of Chinese textbooks to provide real-time assistance to teachers who use the system to adapt Chinese texts.

Finally, with the increasing availability of learner corpora, i.e., corpora of language samples produced by second or foreign language learners, investigations of patterns of as well as individual variation in learner language use, acquisition, and development have flourished. Findings from such investigations have clear values for helping teachers understand learners' abilities, challenges, and developmental trajectories. A good example of this line of research in the current volume is Xu et al. analysis of Chinese learners' acquisition of the Chinese particle *le* using data from the Guangwai-Lancaster Chinese Learner Corpus.

3 Computational Linguistics, Natural Language Processing, and Chinese Language Learning

Computational linguistics, also referred to as natural language processing, is a scientific discipline under the larger field of artificial intelligence that works on problems involving human languages from a computational perceptive (Manning and Schütze 1999). The general objective of computational linguistics is to use computers to construct formal models, either rule-based or statistical, to explain different kinds of linguistic phenomena and to automatically perform tasks that involve the processing, understanding, and/or generation of natural language. Notably, the last two decades have seen a surge of research in developing statistical modeling paradigms from researchers and practitioners in the computational linguistics community, due largely to the explosion of available text and speech data, amid the rapid advancement of machine learning techniques and exponential growth in computing power. These statistical modeling paradigms have been used in many academic prototype and practical product systems, providing such wide-ranging functionalities as syntactic and semantic analysis of text content, automatic speech recognition, information retrieval,

document summarization, machine translation, and question answering, among others. For example, various forms of hidden Markov models (HMM) have been applied to part-of-speech (POS) tagging and speech recognition (Gales and Young 2007), serving as simple and effective frameworks for modeling time-evolving word-level tag sequences and speech vector sequences, respectively. As another illustration, latent Dirichlet allocation (LDA) (Blei et al. 2003) and its two prominent precursors, namely latent semantic analysis (LSA) (Deerwester et al. 1990) and probabilistic latent semantic analysis (PLSA) (Hofmann 1999), have been put to good use in semantic analysis of text content, query-document relevance assessment in information retrieval, and salient sentence selection in extractive document summarization. More recently, deep learning, embodied also as disparate deep neural networks (DNN) and representation learning techniques (Goodfellow et al. 2016), has emerged as the dominant statistical modeling paradigm. The record-breaking performance achieved by this paradigm on a vast array of speech and language processing tasks has exerted a catalytic effect on reviving interest in computational linguistics. One attractive advantage of deep learning is that the associated statistical models can be trained in an end-to-end manner, alleviating the need for specific domain knowledge or human-engineered features while being easily modifiable to accommodate new domains.

Numerous state-of-the-art theories and models of computational linguistics have been applied with good success to computer-assisted language teaching, learning, and assessment, notably in the context of Chinese as a second or foreign language. For example, near-synonym sets, which represent groups of words with similar meanings, are useful knowledge resources for computer-assisted language learning. Recent work has exploited LSA as a context analysis technique to identify useful latent context features for near-synonyms through indirect associations between words and sentences, represented by low-dimensional latent vectors. Classifiers trained on top of such latent vectors can then be used to distinguish among near-synonyms (Wang and Hirst 2010). As a second illustration, computer-assisted pronunciation training (CAPT) systems that are able to pinpoint and diagnose erroneous pronunciations in the utterances produced by L2 learners in response to given text prompts can serve as a surrogate for or supplement to classroom instruction by teachers. A common practice for mispronunciation detection is to extract decision features (attributes) from the prediction output of phone-level acoustic models, which are normally estimated based on certain criteria that maximize the automatic speech recognition performance. The extracted decision features are then fed into a pretrained classifier (decision function) for identifying mispronounced speech segments (Chen and Li 2016). In addition, there have been many attempts to employ long short-term memory (LSTM), a particular type of recurrent neural network (RNN), to detect inappropriate linguistic usage, including character-level confusions (commonly known as spelling errors) and word-, sentence-, or discourse-level grammatical errors (Yeh et al. 2016). Others still, for the task of automated essay scoring (AES), various linguistic features that align with the criteria specified in rating rubrics (e.g., lexical diversity, syntactic complexity, coherence, among others) have been used as predictors of writing quality. Alternatively, some approaches to AES employ LSA to convert a given essay

into a semantic vector (Peng et al. 2010), which represents the location of the essay in a latent semantic space. Close proximity of the semantic vectors representing different essays then implies a high degree of similarity of the semantic content of the essays. Several chapters in this volume detail recent developments in these and other educational NLP applications for Chinese language teaching, learning, and assessment.

4 Organization of the Book

This book is thematically organized into four parts. Part 1, "Introduction," includes two additional introductory chapters beyond the current overview chapter. The second chapter in this part, "Corpus and Computational Methods for Usage-Based Chinese Language Learning: Toward a Professional Multilingualism," starts with a discussion of the main tenets of the usage-based model of language, the theoretical model underlying much of the research presented in this volume. It then illustrates the necessity of corpus and computational methods as tools for usage-based second language acquisition research and the criticality of the voice such methods constitute in the professional multilingualism that informs Chinese language teaching, learning, and assessment.

Xu's chapter, "The Corpus Approach to the Teaching and Learning of Chinese as an L1 and an L2 in Retrospect," reviews the historical development of research into the use of corpora in teaching and learning Chinese. This chapter synthesizes the insights from corpus-based Chinese studies into Chinese textbook compilation and syllabus design, the range and types of important vocabulary items and grammatical patterns to be taught, and the order in which such items and patterns are best to be presented. It further illustrates how some of these insights have found their way into Chinese language pedagogy in the form of data-driven learning and calls for additional research in transforming such insights into pedagogical practices.

Part 2, "Tools, Resources and General Applications," includes three chapters that each introduces a specific type of corpus resource or computational tool with broad pedagogical applications. The first chapter in this part, "Academic Chinese: From Corpora to Language Teaching," describes two written academic Chinese corpora constructed by researchers at National Taiwan Normal University and University of California Los Angeles, respectively. The chapter also reports the outcomes of research on developing lists of academic vocabulary and multiword expressions based on these corpora and discusses applications of the corpora and related research outcomes in teaching material design and pedagogical practices.

Bluemel's chapter, "Pedagogical Applications of Chinese Parallel Corpora," starts with a thorough discussion of the ways in which parallel corpora can be used in language pedagogy in general and then takes a close look at the application of a Chinese-English parallel corpus in a beginning-level Chinese as a foreign language course for L1 English learners. The chapter also includes a review of various parallel corpus resources along with recommendations for their pedagogical uses.

Bo et al.'s chapter, "Data-Driven Adapting for Fine-Tuning Chinese Teaching Materials: Using Corpora as Benchmarks," examines how teachers adapt Chinese texts using an online data-driven text adaptation system, *Chi-Editor*. The system is designed to automatically assess text complexity, to annotate a text for difficult words and long sentences based on features mined from a large textbook corpus, and to dynamically reassess the complexity of the text as it is being modified. The chapter reports evidence that teachers can benefit from the data-driven text adaptation system and discusses how it can be integrated in regular teaching material preparation.

Part 3, "Specific Applications," includes three chapters that each addresses a specific application of a computational tool or corpus resource in Chinese language teaching and learning. The first chapter in this part, "Context Analysis for Computer-Assisted Near-Synonym Learning," reports on the design and performance of a prototype computer-assisted near-synonym learning system for Chinese and English. The system integrates a number of context analysis techniques that can be used to automate the selection of near-synonyms in a given context and to provide useful contextual information to help learners understand different usages of near-synonyms.

Zhang's chapter, "Visualizing Stylistic Differences in Chinese Synonyms," demonstrates the use of Correspondence Analysis to visualize stylistic differences in near-synonyms as two-dimensional bi-plots. The pedagogical advantages of such an approach in promoting Chinese learners' stylistic awareness are showcased using five groups of Chinese near-synonyms.

Gong and Hong's chapter, "Using Corpus-Based Analysis of Neologisms on China's New Media for Teaching Chinese as a Second or Foreign Language," explicates how Chinese neologisms on new media can constitute a productive resource for helping learners acquire and reinforce lexical and sociocultural knowledge. The pedagogical value of information about word-formation strategies of Chinese neologisms is demonstrated through a detailed analysis of a small corpus of authentic usage samples of 50 neologisms originated from Chinese new media.

The fourth part, "Learner Language Analysis and Assessment," includes four chapters that report research on the use of computational and corpus methods for analyzing or assessing learner production. The first chapter in this part, "Acquisition of the Chinese Particle *le* by L2 Learners: A Corpus-Based Approach," uses the spoken subcorpus of the large-scale Guangwai-Lancaster Chinese Learner Corpus to investigate the frequency and accuracy with which Chinese learners at different proficiency levels use the Chinese particle *le* and the developmental pattern of their acquisition of this particle.

Chen and Hsu's chapter, "Mandarin Chinese Mispronunciation Detection and Diagnosis Leveraging Deep Neural Network based Acoustic Modeling and Training Techniques," focuses on automatic mispronunciation detection and diagnosis. Following a review of recent developments on this front with state-of-the-art automatic speech recognition methodologies, the chapter presents a training approach for deep neural network based acoustic models as well as a series of empirical experiments designed to evaluate the effectiveness of the approach.

Lee et al.'s chapter, "Resources and Evaluations of Automated Chinese Error Diagnosis for Language Learners," introduces the HSK Dynamic Composition Cor-

pus and the TOCFL Learner Corpus as useful resources for understanding patterns of Chinese learners' grammatical errors as well as a series of bakeoffs and shared tasks that utilized these corpora to evaluate computational systems for Chinese spelling checking and grammatical error identification.

The final chapter, "Automated Chinese Essay Scoring Based on Multilevel Linguistic Features," discusses critical issues concerning the task of automated Chinese essay scoring, followed by a detailed description of the design of a novel automated system for scoring Chinese essays along with experimental results indicating superior stability and reliability of its performance.

References

Aijmer, K. (Ed.). (2009). *Corpora and language teaching*. Amsterdam/Philadelphia: John Benjamins.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(4–5), 993–1022.

Chen, N. F., & Li, H. (2016). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–417.

Díaz-Negrillo, A., Ballier, N., & Thompson, P. (Eds.). (2013). *Automatic treatment and analysis of learner corpus data* (pp. 249–264). Amsterdam/Philadelphia: John Benjamins.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy. *International Journal of Corpus Linguistics, 14*(3), 393–417.

Gales, M., & Young, S. (2007). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing, 3*(1), 195–304.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: The MIT Press.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50–57).

Jin, T., & Lu, X. (2018). A data-driven approach to text adaptation in teaching material preparation: Design, implementation and teacher professional development. *TESOL Quarterly, 52*(2), 457–467.

Jurafsky, D., & Martin, J. (2008). *An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Dordrecht: Springer.

Lu, X. (2018). Natural language processing and intelligent computer-assisted language learning (ICALL). In J. I. Liontas (Ed.), *The TESOL encyclopedia of english language teaching*. Chichester, UK: Wiley Blackwell.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Peng, X., Ke, D., Chen, Z., & Xu, B. (2010). Automated Chinese essay scoring using vector space models. In *Proceedings of the International Universal Communication Symposium* (pp. 149–153).

Wang, T., & Hirst, G. (2010). Near-synonym lexical choice in latent semantic space. In *Proceedings of the International Conference on Computational Linguistics* (pp. 1182–1190).

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. London: Routledge.

Yeh, J.-F., Hsu, T.-W., & Yeh, C.-K. (2016). Grammatical error detection based on machine learning for Mandarin as second language learning. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 140–147).