The Jinan Chinese Learner Corpus

Maolin Wang[1], **Shervin Malmasi**[2] **and Mingxuan Huang**[3]

[1]College of Chinese Language and Culture, Jinan University, Guangzhou, China
[2]Centre for Language Technology, Macquarie University, Sydney, NSW, Australia
[3]Guangxi University of Finance and Economics, Nanning, China

[1]wmljd@126.com, [2]shervin.malmasi@mq.edu.au, [3]gxmingxh@163.com

Abstract

We present the Jinan Chinese Learner Corpus, a large collection of L2 Chinese texts produced by learners that can be used for educational tasks. The present work introduces the data and provides a detailed description. Currently, the corpus contains approximately 6 million Chinese characters written by students from over 50 different L1 backgrounds. This is a large-scale corpus of learner Chinese texts which is freely available to researchers either through a web interface or as a set of raw texts. The data can be used in NLP tasks including automatic essay grading, language transfer analysis and error detection and correction. It can also be used in applied and corpus linguistics to support Second Language Acquisition (SLA) research and the development of pedagogical resources. Practical applications of the data and future directions are discussed.

1 Introduction

Despite the rapid growth of learner corpus research in recent years (Díaz-Negrillo et al., 2013), no large-scale corpus of second language (L2) Chinese has been made readily available to the research community.

Learner corpora are often used to investigate learner language production in an exploratory manner in order to generate hypotheses about learner language. Recently, learner corpora have also been utilized in various educational NLP tasks including error detection and correction (Gamon et al., 2013), Native Language Identification (Tetreault et al., 2013) and language transfer hypothesis formulation (Swanson and Charniak, 2014).

While such corpus-based studies have become an accepted standard in SLA research and relevant NLP tasks, there remains a paucity of large-scale L2 corpora. For L2 English, the two main datasets are the ICLE (Granger, 2003) and TOEFL11 (Blanchard et al., 2013) corpora, with the latter being the largest publicly available corpus of non-native English writing.[1] However, this data scarcity is far more acute for L2 other than English and this has not gone unnoticed by the research community (Lozano and Mendikoetxea, 2013; Abuhakema et al., 2008).

The present work attempts to address this gap by making available the Jinan Chinese Learner Corpus (JCLC), an L2 Chinese corpus designed for use in NLP, corpus linguistics and other educational domains. This corpus stands out for its considerable size and breadth of data collection. Furthermore, the corpus – an ongoing project since 2006 – continues to be expanded with new data. In releasing this data we hope to equip researchers with the data to support numerous research directions[2] going forward.

The JCLC is freely available to the research community and accessible via our website.[3] It can be used via a web-based interface for querying the data. Alternatively, the original texts can be downloaded in text format for more advanced tasks.

[1]TOEFL11 contains over 4 million tokens in 12,100 texts.
[2]See section 5 for examples.
[3]http://hwy.jnu.edu.cn/jclc/

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 118–123,
Denver, Colorado, June 4, 2015.

2 Background

Interest in learning Chinese is rapidly growing, leading to increased research in Teaching Chinese as a Foreign Language (TCFL) and the development of related resources such as learner corpora (Chen et al., 2010).

This booming growth in Chinese language learning (Rose and Carson, 2014; Zhao and Huang, 2010), related to the dramatic globalization of the past few decades and a shift in the global language order (Tsung and Cruickshank, 2011), has brought with it learners from diverse backgrounds. Consequently, a key challenge here is the development of appropriate resources – language learning tools, assessments and pedagogical materials – driven by language technology, applied linguistics and SLA research (Tsung and Cruickshank, 2011). The application of these tools and SLA research can greatly assist researchers in creating effective teaching practices and is an area of active research.

This pattern of growing interest in Chinese is also reflected in the NLP community, evidenced by the continuously increasing research focus on Chinese tools and resources (Wong et al., 2009).

A key application of such corpora is in the field of Second Language Acquisition (SLA) which aims to build models of language acquisition. One aspect of SLA is to formulate and test hypotheses about particularly common patterns of difficulty that impede L2 production among students. This is usually done using the natural language produced by learners to identify deficits in their interlanguage.

A criticism of SLA has been that its empirical foundation is weak (Granger, 2002), casting doubts on the generalizability of results. However, this is beginning to change with the shift towards using large learner corpora. The creation of such corpora has led to an efflorescence of empirical research into language acquisition (Granger, 2002).

The use of NLP and machine learning methods has also extended to SLA, with a new focus on a combined multidisciplinary approach to developing methods for extracting ranked lists of language transfer candidates (Swanson and Charniak, 2014; Malmasi and Dras, 2014c).

3 Data Collection and Design

The JCLC project, started in 2006, aims to create a corpus of non-native Chinese texts, similar to the ICLE. The majority of the data has been collected from foreign students learning Chinese at various universities in China, with some data coming from universities outside China. This data includes both exams and assignments. The texts are manually transcribed with all errors being maintained. Error annotations are not available at this stage.

In order to be representative, the corpus includes student data from a wide range of countries and proficiency levels. 59 different nationalities are represented in the corpus. Proficiency levels are classified according to the length of study and include: beginners (less than 1 year), intermediate (2-3 years) and advanced (3+ years). In selecting texts for inclusion, we strived to maximize representativeness across all proficiencies.

3.1 Data Format

The learner texts are made available as Unicode (UTF-8) text files to ensure maximum compatibility with linguistic and NLP tools.

3.2 Metadata

In order to support different research directions, extensive metadata about each text has been recorded. This metadata is available in text, CSV and Microsoft Excel format. The variables are outlined below.

Writing ID A unique id assigned to each text.

Writing Type Either exam or assignment.

Student ID While student names are redacted, they are each assigned a unique ID which allows for the analysis of longitudinal data in the corpus.

Date The submission date of the writing also enables longitudinal analysis of a student's data.

Gender, Age and Education level This data allows the investigation of other research questions, e.g. the critical age hypothesis (Birdsong, 1999).

Native Language This variable is helpful in studying language transfer effects by taking into account the author's native language.

Other Acquired Languages It should be noted that the currently used learner corpora, including the ICLE and TOEFL11, fail to distinguish whether the learner language is in fact the writer's second language, or if it is possibly a third language (L3). It has been noted in the SLA literature that when acquiring an L3, there may be instances of both L1- and L2-based transfer effects on L3 production (Ringbom, 2001). Studies of such second language transfer effects during third language acquisition have been a recent focus on cross-linguistic influence research (Murphy, 2005). The JCLC is the first large-scale learner corpus to include this information as well.

Proficiency Level Determined by the length of study, as described above, a level of beginner, intermediate or advanced is assigned to each text.

Length of Chinese study The amount of time spent studying Chinese. Study inside and outside China are recorded separately.

Chinese heritage learner This variable indicates if the learner is of Chinese heritage, was exposed to Chinese at home, and if so, which dialect.

4 Corpus Analysis

We now turn to a brief analysis of the corpus. The current version of the JCLC contains 5.91 million Chinese characters across 8,739 texts. The top backgrounds of the learners and their text frequency and mean lengths[4] are shown in Table 1.

We also observe high variability in text lengths across the data.[5] A histogram of the text lengths, shown in Figure 1, confirms this trend. We believe that this is a result of the data being collected from a variety of tasks of different scopes from a range of courses at different institutes. Most texts fall in the 250-700 token range of the distribution.

For text types, 57% of the texts are assignments while the remaining 43% are mostly exams.

We can also look at the distribution of proficiency levels in the data, as shown in Figure 2. The majority of the texts, 65%, fall into the medium category with 21% and 14% in the low and high levels, respectively. Comparing this distribution to that of the data in the TOEFL11, also shown in Figure 2, we observe a similar trend with the great majority of the data falling into the medium proficiency bracket. The TOEFL11 has more advanced learners, which is to be expected given that the texts are all collected from a high-stakes exam.

While the data sampling is not equal across all language/proficiency groups we note that this type of imbalance is a perennial problem present in most learner corpora and generally a result of the demographics of the students. Given these constraints, we strived to adhere to key corpus design principles (Wynne, 2005) at all stages.

Language	Texts	Mean Token Count
Indonesian	3381	663.62
Thai	1307	755.86
Vietnamese	824	721.41
Korean	568	399.45
Burmese	410	776.92
Laotian	398	794.78
Khmer	329	691.62
Filipino	293	1135.90
Japanese	270	446.13
Spanish	198	401.85
Mongolian	119	537.02
Others	642	418.26
Total	**8739**	**675.93**

Table 1: The top native language backgrounds available in the corpus, including document counts and the average number of Chinese tokens per text.

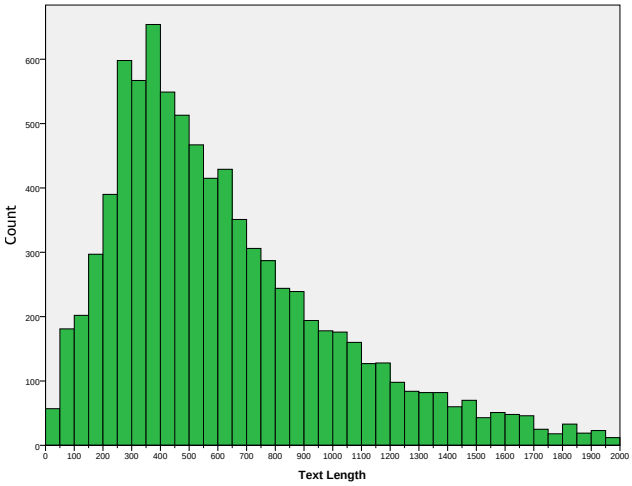

Figure 1: Histogram of text lengths (bin size = 50).

[4] As measured by the number of Chinese characters.

[5] The standard deviation in text length is 530 tokens.

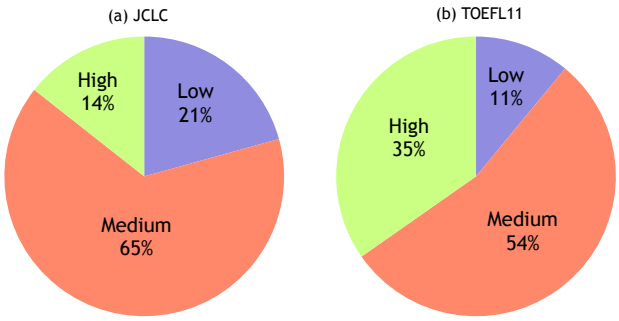

Figure 2: Proficiency distributions in the Jinan Chinese Learner Corpus (left) and the TOEFL11 corpus (right).

In sum, we see that the JCLC is a large corpus and represents various native language and proficiency groups. These characteristics make it suitable for a wide range of research tasks, as described in the next section.

5 Applications

Educational studies in linguistics and NLP have been increasing recently. To this end, this corpus can be used in various areas, as outlined here.

Automatic Essay Scoring is an active area of research that relies on examining the differences between proficiency levels using large learner data and NLP methods (Yannakoudakis et al., 2011). Given the inclusion of proficiency data, the JCLC could also be used to investigate the extension of current automatic grading techniques to Chinese, something which has not been done to date.

Error Detection and Correction There is growing research in building error detection and correction systems trained on learner corpus data (Dahlmeier and Ng, 2011; Han et al., 2010). This was also the focus of a recent shared tasks including Helping Our Own (Dale and Kilgarriff, 2011; Dale et al., 2012) and CoNLL shared tasks (Ng et al., 2013). A recent shared task also focused on Chinese error correction (Yu et al., 2014). This research was also recently extended to Chinese word ordering error detection and correction (Cheng et al., 2014), also using learner texts. The large JCLC can be used in such tasks through the addition of error annotations.

Native Language Identification is the task of inferring an author's native tongue based on their writings in another language (Malmasi and Dras, 2015). This task mainly relies on learner corpora and the JCLC could be directly applied here. A good overview is presented in the review of the recent NLI shared task (Tetreault et al., 2013). NLI methods have already been tested on other languages including Arabic and Finnish (Malmasi and Dras, 2014a; Malmasi and Dras, 2014b).

Transfer Hypothesis Extraction Researchers have recently investigated using data-driven techniques combined with machine learning and NLP to extract language transfer hypotheses from learner corpora (Swanson and Charniak, 2014).

Second Language Acquisition researchers are interested in contrasting the productions of natives and non-natives (Housen, 2002). This is made possible with the JCLC data and the presence of multiple L1s allows for contrastive interlanguage analysis between different native languages as well. The availability of such large-scale data with different L1-L2 combinations can enable broad language acquisition research that can be extrapolated to other learners.

Pedagogical Material Development Learner corpora have been used identify areas of difficulty and enable material designers to create resources that take into account the strengths and weaknesses of students from distinct groups (McEnery and Xiao, 2011). This can also be further expanded to syllabus development where corpus-derived knowledge can be used to guide the design process.

Combined with language transfer analysis, learner data can be used to aid development of pedagogical material within a needs-based and data-driven approach. Once language use patterns are uncovered, they can be assessed for teachability and used to create tailored, native language-specific exercises and teaching material.

Automatic Assessment Generation Combined with the above-mentioned error detection and language transfer extraction methods, this data can be used to automatically generate testing material (e.g. Cloze tests). Following such an approach, recent work by Sakaguchi et al. (2013) made use of large-scale English learner data to generate fill-in-the-

blank quiz items for language learners. Previous research in this space had also considered the automatic generation of multiple-choice questions for language testing (Hoshino and Nakagawa, 2005), but without learner data. The use of learner corpora containing naturally produced errors provides a much more promising synergy, enabling the assessment of more complex linguistic errors beyond articles, prepositions and synonyms. With further annotations of the present errors, the JCLC could be used for such tasks.

6 Conclusion and Future Work

The JCLC, a sizeable project that has been ongoing for the last 8 years, has yielded a large-scale language resource for researchers – the first of its kind. As the only such corpus of this size, the JCLC is a valuable resource to support research in various areas, some of which we outlined here.

Research in most of the tasks described in section 5 has focused on English. The availability of the JCLC will enable much of this work to be extended to Chinese, potentially opening new research areas for the community.

The JCLC is an ongoing project and new data continues to be collected and added to the corpus. No fixed target size has been set and it is anticipated that the corpus will grow to be much larger than the current size.

Several directions for future work are under consideration. One avenue is the the creation of further annotation layers over the data to include additional linguistic information such as Chinese word segmentation boundaries, part-of-speech tags, constituency parses and grammatical dependencies. The inclusion of error annotations and manual corrections is another potential avenue for future work.

Another possibility is the addition of a new subcorpus of native texts that can be used as a control group for comparing native and non-native data. This would enable further analysis of learner interlanguage.

Acknowledgments

We would like to thank the three anonymous reviewers for their insightful comments.

References

Ghazi Abuhakema, Reem Faraj, Anna Feldman, and Eileen Fitzpatrick. 2008. Annotating an Arabic Learner Corpus for Error. In *LREC*.

David Birdsong. 1999. *Second Language Acquisition and the Critical Period Hypothesis*. Routledge.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service.

Jianguo Chen, Chuang Wang, and Jinfa Cai. 2010. *Teaching and learning Chinese: Issues and perspectives*. IAP.

Shuk-Man Cheng, Chi-Hsin Yu, and Hsin-Hsi Chen. 2014. Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. *Proceedings of COLING 2014*, pages 279–289.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 915–923. Association for Computational Linguistics.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249. Association for Computational Linguistics.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62. Association for Computational Linguistics.

Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson. 2013. *Automatic treatment and analysis of learner corpus data*, volume 59. John Benjamins Publishing Company.

Michael Gamon, Martin Chodorow, Claudia Leacock, Joel Tetreault, N Ballier, A Díaz-Negrillo, and P Thompson. 2013. Using learner corpora for automatic error detection and correction. *Automatic treatment and analysis of learner corpus data*, pages 127–150.

Sylviane Granger. 2002. A bird's-eye view of learner corpus research. *Computer learner corpora, second language acquisition and foreign language teaching*, pages 3–33.

Sylviane Granger. 2003. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3):538–546.

Na-Rae Han, Joel R Tetreault, Soo-Hwa Lee, and Jin-Young Ha. 2010. Using an error-annotated learner corpus to develop an esl/efl error correction system. In *LREC*.

Ayako Hoshino and Hiroshi Nakagawa. 2005. A real-time multiple-choice question generation for language testing: a preliminary study. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 17–20. Association for Computational Linguistics.

Alex Housen. 2002. A corpus-based study of the L2-acquisition of the English verb system. *Computer learner corpora, second language acquisition and foreign language teaching*, 6:2002–77.

Cristóbal Lozano and Amaya Mendikoetxea. 2013. Learner corpora and second language acquisition. *Automatic Treatment and Analysis of Learner Corpus Data*, 59.

Shervin Malmasi and Mark Dras. 2014a. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (EMNLP 2014)*, pages 180–186, Doha, Qatar, October. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014b. Finnish Native Language Identification. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 139–144, Melbourne, Australia.

Shervin Malmasi and Mark Dras. 2014c. Language Transfer Hypotheses with Linear SVM Weights. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shervin Malmasi and Mark Dras. 2015. Large-scale Native Language Identification with Cross-Corpus Evaluation. In *Proceedings of NAACL-HLT 2015*, Denver, Colorado, June. Association for Computational Linguistics.

Tony McEnery and Richard Xiao. 2011. What corpora can offer in language teaching and learning. *Handbook of research in second language teaching and learning. London: Routledge*, pages 364–380.

Shirin Murphy. 2005. Second language transfer during third language acquisition. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 3(1).

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of CoNLL*.

Hakan Ringbom. 2001. Lexical transfer in L3 production. volume 31, pages 59–68. Multilingual Matters.

Heath Rose and Lorna Carson. 2014. Introduction. *Language Learning in Higher Education*, 4(2):257–269.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 238–242.

Ben Swanson and Eugene Charniak. 2014. Data Driven Language Transfer Hypotheses. *EACL 2014*, page 169.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A Report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia, June. Association for Computational Linguistics.

Linda Tsung and Ken Cruickshank. 2011. *Teaching and Learning Chinese in Global Contexts: CFL Worldwide*. Bloomsbury Publishing.

Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zhengsheng Zhang. 2009. Introduction to Chinese Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 2(1):1–148.

Martin Wynne, editor. 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxbow Books Oxford.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of Grammatical Error Diagnosis for Learning Chinese as a Foreign Language. In *Proceedings of the 22nd International Conference on Computers in Education*, Nara, Japan.

Hongqin Zhao and Jianbin Huang. 2010. Chinas policy of Chinese as a foreign language and the use of overseas Confucius Institutes. *Educational Research for Policy and Practice*, 9(2):127–142.