

Parallel Corpora in Pedagogy: Using Parallel Corpora in Teaching Chinese as a Foreign Language

Brody Bluemel
Ph.D. candidate
Department of Applied Linguistics
The Pennsylvania State University
btb5129@psu.edu

Abstract

This report provides a brief outline of the project completed for my dissertation research. The project included the development of a web-based Chinese/English parallel corpus, *www.parallelcorpus.com*, which was designed as a learning tool to assist English L1 learners of Chinese in the acquisition of written Chinese. This corpus tool was then implemented in two beginning-level high school Chinese classrooms over a period of three months, providing evidence of positive learning outcomes and suggesting the potential for further pedagogical applications of parallel corpus technology.

Introduction

As technology has advanced, corpora have evolved into a valuable tool in language research and pedagogy. The growth witnessed in corpus research, design, and application has led to the further development of this technology and to the formation of many new specialized corpora. Parallel corpora are specialized translation corpora that consist of source texts in one language that are aligned with translated texts in one or more additional languages. This corpus technology has been in use for the past twenty years in comparative linguistics and translation studies, but has only recently been recognized for its value in foreign language pedagogy. The limited classroom studies and applications of parallel corpora have affirmed their worth, but also unite in the call for further exploration of this technology.

The premise of the current study responds to this call through the development of a new web-based parallel corpus tool designed specifically for pedagogical application [*www.parallelcorpus.com*]. This corpus tool was designed with the intent of capitalizing on the unique features of parallel corpora in order to more effectively assist language learners in the development of reading, writing, and lexical acquisition in their target language. The initial application of this tool was created for English L1 learners of Chinese. Chinese has a uniquely complex orthography that is a recognized challenge for both native and foreign language learners. The features of *www.parallelcorpus.com* seek to make this complex writing system more accessible to learners through the use of the transliterated *pinyin* system, and the use of the students' native language.

The parallel corpus tool was first applied in two beginning-level high school Chinese classrooms over a three-month period. The students used the tool both in-class and outside of class to complete a series of assignments, projects, and exams. An overall positive learning outcome was observed, with the parallel corpus tool functioning as an effective medium in written language acquisition.

The Parallel Corpus Tool

Parallel corpus technology is still developing, and there are a limited number of parallel corpora available for use. Additionally, of the corpora that are available, the majority of them were designed for linguistics research and not for pedagogical application. The corpus tool designed for this study – www.parallelcorpus.com - was developed specifically for pedagogical application and addressed issues students face in learning the Chinese orthographic system. Specifically, it sought to address the problems of Chinese character acquisition and tone recognition.

To understand these two issues, and how the parallel corpus addresses them, it is first necessary to outline a few key features of the Chinese language. Chinese is written with characters, but characters are also represented through *pinyin*, a phonetic system for transliterating Chinese into the Roman alphabet. Each individual character corresponds to one syllable that also has a specific tonal pronunciation. In Mandarin Chinese there are four tones, plus an additional no-tone pronunciation, thus there are five different potential tonal pronunciations for each individual character. The pinyin system makes use of diacritic marks, known as tone marks, to indicate the correct tone for a particular character. The challenge for a student lies in the fact that the pinyin representation and tonal pronunciation are not encoded within the characters. Therefore the tone and pinyin have to be learned for each individual character. Given that one must know an estimated 3,500 characters to read an average Chinese text, the task seems even greater. The traditional approach to mastering Chinese orthography has heavily emphasized memorization. Parallel corpus technology offers a functional alternative that effectively provides a method for electronically encoding both pinyin and tone in each individual character, making it possible for even beginning-level students to work through Chinese texts.

The tool accomplishes this by expanding the general functions and features of parallel corpora. Instead of presenting the texts in a bilingual Chinese/English layout, four different language formats are provided: Chinese characters, characters + tone marks, pinyin, and English. The corpus texts are aligned in all four formats, but the user is able to control which of the formats are visible at any given time (Figure 1). The Chinese text format is interactive allowing the user to click on any individual character once to view its tone diacritic and a second time to view the pinyin. Effectively, the two additional language formats – characters + tone marks and pinyin – function as a form of interlanguage to aid learners in mastering Chinese characters. While the pinyin representation of characters is available in most Chinese language technologies, the implementation of the character + tone format is a new approach, which to my knowledge, has not been used previously. Figure 1 shows a screen caption of an example in the parallel corpus tool.

Figure 1

Search Results: 高兴			
	Chinese	Pinyin	English
13 results	高兴		
China	Pinyin Text	English text	
2. 他很高兴他收到很多礼物	2. Tā hěn gāoxìng tā shōudào hěnduō lǐwù	2. He is very happy he received a lot of presents	
3. 她很高兴她收到很多压岁钱	3. Tā hěn gāoxìng tā shōudào hěnduō yāsuiqián	3. She is very happy she received a lot of spring festival money.	
很高兴看到你们齐心协力干活!	Hěn gāoxìng kàndào nǐmen qíxīn héli gàn huó!	It is good to see you working together!	More...

The purpose behind the interactive nature of the corpus tool is to enable the user to self-regulate their reading experience by using the least amount of mediation necessary to read and comprehend a text. Ideally, a learner would be able to use just the character version, but the other formats make it possible for a user to retrieve specific elements they are unsure of whether it be tone, pronunciation or meaning. All of the interactive features of the tool are accessible whether the user is viewing a complete sample text or if they are querying a specific term or phrase.

The method used in compiling texts for this parallel corpus also varies from traditional approaches. The corpus is comprised of bidirectional (Chinese L1-English L2 translations *and* English L1- Chinese L2 translations) that are functionally accessible to students. A basic set of texts including the students' textbook and additional materials that address the topics covered in the class were included in the initial subset of texts accessible to students in the corpus tool. The course instructor, however, had control over the data visible to the students – meaning that the instructor could hide from view any of the existing texts, and/or add additional material at any given time. The purpose of providing the instructor with this control was not to limit students' exposure to texts, but rather to insure the students are able to access contextually appropriate material that they could manage, but also be challenged by. This feature also allows the instructor to add more difficult texts throughout the term as the students continue to learn and develop.

Classroom Application and Methodology

The parallel corpus tool was implemented simultaneously in two different beginning-level high school Chinese classes. The classes were taught by the same instructor/researcher for the entire academic year, but the study was conducted over a three-month period toward the end of the school year. The students were all high school students (grades 9-12), and had been studying Chinese for approximately five months when the study began. Both classes followed identical implementation procedures. During the first month of the study no changes were made to the class, but data regarding student performance

and reflections was collected. In month two, *www.parallelcorpus.com* was introduced and used regularly both by the instructor and students in the class, and students also had open access to the tool outside of the classroom. Students were encouraged to use the tool during this time, but not required. By the beginning of the final month of the study, the tool was fully integrated in the classroom and students were required to use the tool in completing homework, in-class assignments, exams, and other activities.

The three different month-long phases of the study were designed to observe how learning took place before the parallel corpus tool was introduced, to follow through the process of implementing the tool, and then to observe the effects of the tool on learner development once it was fully integrated in the classroom. The second month of the study also allowed time for students to become familiar and comfortable with using this new technology, before it became a required element of their learning experience.

In order to observe how the tool was used, and the effect it had on student learning several different sources of data were collected. First, all written tasks completed by students throughout the duration of the study were collected. This included all homework assignments, in-class assignments, course projects, and exams. These materials were later analyzed to evaluate students' performance in reading comprehension, writing, and lexical acquisition tasks. Second, screen-recordings were taken of students' computer monitors during all in-class tasks and exams in which *www.parallelcorpus.com* was used. These recordings depict exactly *how* the students chose to use the tool, and what features they relied on the most. Though the students were encouraged and/or required to complete different tasks and exams with the tool, they were not restricted or instructed in *how* to use the tool, thus the screen-recordings provided insight into what features the students found most beneficial, and which they ignored. Third, the students were required to complete learner logs with every course assignment and exam. This was part of the regular curriculum of the class and intended to help students become aware of how they were studying and learning the language and what tools and approaches worked best for them individually. The learner logs detailed what resources students employed in completing a task and also, through self-assessments, shed light onto the challenges students faced or insights they gained. Fourth, an end of semester questionnaire was completed, which required the students to evaluate and provided explicit feedback on their experience using the parallel corpus. Finally, the instructor/researcher maintained an auto-ethnography during the three-month study that detailed the implementation of the tool.

The rich and varied sources of data make it possible to analyze several key elements of implementing the corpus tool in a Chinese classroom. It allows for a general assessment of the effectiveness of the tool in assisting learners in their development of reading comprehension, writing, and lexical acquisition. Additionally, it provides insight into how to best apply this and similar technologies in the classroom. This is accomplished both by considering the student assessments and perspectives provided through the learner logs and questionnaire, as well as by evaluating the instructor/researcher's auto-ethnography of the successes and challenges faced during the process.

Conclusion

By developing a new parallel corpus tool, this project sought to both provide an innovative approach to learning written Chinese and to explore a new pedagogical application of parallel corpora. Preliminary analysis of the results of the project indicate that the tool was very effective in aiding students in their study of written Chinese by making the language more functionally accessible to them. Further, implementing the tool also provided insights into how parallel corpora can be utilized in foreign language learning, as well as suggesting how this technology can be further developed to address unique challenges associated with various individual languages. More detailed results will be made available through future publications and interested readers are invited to follow the project through announcements on the Corpus Portal website of the Center for Advanced Language Proficiency Education and Research.

*Please refer to this text as:
Bluemel, B. (2013). Parallel Corpora in Pedagogy: Using Parallel Corpora in Teaching Chinese as a Foreign Language (CALPER Corpus Community Report, No. 5). The Pennsylvania State University: Center for Advanced Language Proficiency Education and Research.*