CrossMark

ReCALL **28**(2): 187–206. 2016 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344015000257
First published online 17 March 2016

Corpus-based learning of Cantonese for Mandarin speakers

TAK-SUM WONG

City University of Hong Kong, Hong Kong
(email: tswong-c@my.cityu.edu.hk)

JOHN S. Y. LEE

City University of Hong Kong, Hong Kong
(email: jsylee@cityu.edu.hk)

Abstract

This article presents the first study on using a parallel corpus to teach Cantonese, the variety of Chinese spoken in Hong Kong. We evaluated this approach with Mandarin-speaking undergraduate students at the beginner level. Exploiting their knowledge of Mandarin, a closely related language, the students studied Cantonese with authentic material in a Cantonese-Mandarin parallel corpus, transcribed from television programs. They were given a list of Mandarin words that yield a range of possible Cantonese translations, depending on the linguistic context. Leveraging sentence and word alignments in the parallel corpus, the students independently searched for example sentences to discover these translation equivalents. Experimental results showed that, in both the short- and long-term, this data-driven learning approach helped students improve their knowledge of Cantonese vocabulary. These results suggest the potential of applying parallel corpora at even the beginners' level for other L1-L2 pairs of closely related languages.

Keywords: parallel corpus, parallel concordance, language learning, Cantonese, Mandarin

1 Introduction

The data-driven learning (DDL) approach gives language students direct access to a corpus of example sentences (Johns, 1994). The approach encourages students themselves to discover the patterns of language use by observing these examples, rather than by learning rules from teachers (Johns, 1991). In past applications of DDL, students have consulted text corpora to identify translation equivalents (e.g., Wang, 2001); to edit their own writing (e.g., Gao, 2011; Geist & Hahn, 2012; Tono, Satake & Miura, 2014); to acquire vocabulary (e.g., Huang & Liou, 2007; Poole, 2012); to learn word collocations (e.g., Chan & Liou, 2005); and to study various aspects of grammar (Tian, 2004; Boulton, 2008b; Boulton, 2008c; Boulton, 2009; Chujo, Anthony & Oghigian, 2009; Boulton, 2010).

In these learning tasks, students typically formulate linguistic hypotheses about some aspects of their second language (L2), then confirm or reject them based on evidence

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

observed in the corpus (Nerbonne, 2000; Montero Perez, Paulussen, Macken & Desmet, 2014); they play, in effect, the role of a linguistic researcher or a detective (Johns, 1997: 101). Because it is so demanding on students, DDL or corpus-based language learning has been attempted mostly with intermediate to advanced level students.

This article aims to address the research question: can language learners at the beginner level autonomously use a parallel corpus to improve their L2 proficiency? Although this question has overall been under-investigated (Boulton, 2008a), there is growing evidence that DDL can be effective for students of English at lower levels (e.g., Boulton, 2009; Chujo *et al*., 2009). We contribute further evidence by reporting experimental results on the case of applying a Cantonese-Mandarin parallel corpus to teach Cantonese to beginners. We evaluated this approach with undergraduate students whose first language (L1) is Mandarin Chinese. Although Cantonese and Mandarin are not mutually intelligible, they are closely related varieties of Chinese. By exploiting their knowledge of Mandarin, students independently studied Cantonese with the parallel corpus, which consists of authentic, advanced Cantonese material transcribed from television programs, along with the Mandarin subtitles.

In a study involving 131 students in a real classroom setting, we show that this approach significantly improved the students' knowledge of Cantonese vocabulary with minimal teacher supervision. These results are encouraging for the introduction of corpus material even for beginner learners – not only for other varieties of Chinese, but also for other closely related L1-L2 pairs including, for example, members of the Romance languages.

2 Background

In this section, we first sketch the linguistic background of Cantonese and Mandarin, then discuss the DDL method, and situate our study among past research on corpora-based methods for language learning.

2.1 Cantonese and Mandarin

Cantonese is the variety of Chinese spoken in Hong Kong, while Mandarin, also known as Putonghua, is the dominant variety in mainland China. Spoken by more than 55 million people, Cantonese is the "most widely known and influential variety of Chinese other than Mandarin" (Matthews & Yip, 2011). Since its return to China in 1997, Hong Kong has received a large number of visitors from mainland China to study and work in the city. There has thus been a marked increase in the need to teach Cantonese to the mainland Chinese, virtually all of whom have native proficiency of Mandarin. The interested reader is referred to Zeng (1993) and Ōuyáng (1993) for detailed contrastive analyses between Cantonese and Mandarin. We now give a brief account to provide the necessary background.

Although mutually unintelligible in their spoken form, Cantonese and Mandarin are genetically related, having both developed from Middle Chinese. They share similar writing systems, as well as many cognates. For example, in Figure 1, the Mandarin *xìn* and Cantonese *seun* 'letter' are both written as the character 信; the classifier for 'letter' is *fēng* in Mandarin and *fūng* in Cantonese, and both are written as the character 封. In a small-scale study, Ōuyáng (1993: 23) estimated that 70% of the words of kinship in the two languages are identical in word form. To evaluate a larger sample, we compared the Cantonese and

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Mandarin equivalents (Li, Wong, Leung & Wong, 2016) of the Leipzig-Jakarta list of 100 basic words (Tadmor, Haspelmath & Taylor, 2010: 239–241). Among these 100 pairs of words, 60% have identical written forms, most of which have highly regular phonological correspondence; a further 20% have the same core morpheme.[1] These lexical and orthographic similarities allow Mandarin speakers, even those with no knowledge of Cantonese, to infer the meaning of much of written Cantonese. In this study, we exploit these linguistic properties to introduce DDL to beginner students, who can leverage their Mandarin knowledge to access advanced Cantonese corpus material.

While Cantonese is written in Hong Kong with 'traditional' characters, Mandarin is written in mainland China with 'simplified' characters, a system introduced about 50 years ago. In our experience, within a month of regular exposure to traditional characters, university students from the mainland can usually read most of these characters when placed in context. Getting used to traditional characters, however, is only a start. The differences between the major Chinese varieties have been described as being "at least on the order of the different languages of the Romance family" (Hannas, 1997: 198), or "roughly parallel to English, Dutch, Swedish, and so on among the Germanic group of the Indo-European language family" (Mair, 1991: 3). In a comparative analysis of Cantonese and Mandarin, Ōuyáng noted that although their "grammatical structure is similar in most major respects, the differences are not insignificant" (1993: 274). These differences include the use of modal verbs and predicative adjectives, the expression of epistemicity and comparative construction, the word order in double object constructions (cf. Figure 1), and the system of sentence-final particles, which is significantly more complicated in Cantonese. Aside from grammatical issues, a major source of difficulty for Mandarin speakers is the difference in sense and semantic domain between words with equivalent forms in Mandarin and Cantonese. For example, the Mandarin 等 *děng* 'wait' and 讓 *ràng* 'give way', when used as verbs, have identical Cantonese counterparts, 等 *dáng* 'wait' and 讓 *yeuhng* 'give way'. When *ràng* functions as a preposition meaning 'let', however, it must not be translated as *yeuhng*, but rather as *dáng*. When Mandarin speakers fail to acquire such differences in semantic domain, they are prone to selecting inappropriate Cantonese *faux-amis* words.

The combination of lexical and orthographic similarities on the one hand, and significant differences in semantic domain on the other, makes Mandarin and Cantonese a potentially interesting L1-L2 pair for introducing DDL to beginner learners. In this study, we will investigate whether such learners can autonomously improve their knowledge in Cantonese vocabulary with a Cantonese-Mandarin parallel corpus containing authentic and advanced text.

2.2 Data-driven learning

Boulton (2011) proposes a prototype definition for DDL: "the hands-on use of authentic corpus data (concordances) by advanced, sophisticated foreign or second language learners in higher education for inductive, self-directed language learning of advanced usage." But he recognizes that the boundary of DDL is fuzzy, and so the characteristics given in the

[1] For example, the character 鼻 serves as the common core morpheme for both the Mandarin bízi 鼻子 'nose' and the Cantonese *beih* 鼻 'nose'.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

definition may vary in different DDL instantiations. In this section, we situate our study in relation to other work in terms of several aspects of this definition.

Corpus data. Corpora have been deployed in language learning for many L1-L2 pairs. Those involving Chinese mostly served to teach English L2 to Chinese speakers. Using an English-Chinese parallel concordance, Wang (2001) conducted a study on the Chinese adverb 現在 *xiànzài* 'now', and reported that Chinese students succeeded in recognizing its various English equivalents. Chan and Liou (2005) showed that the use of a web-based Chinese-English concordance can improve knowledge of English verb-noun collocations for university EFL students. Huang and Liou (2007) compiled a corpus of articles from a bilingual Chinese-English magazine, and showed that participants improved their vocabulary scores after a twelve-week reading program. Gao (2011) deployed the CERT parallel Chinese-English concordance to help learners independently improve their English writing. When evaluated on Chinese-to-English translation, participants received significantly better scores, mostly by revising their word choices and word combinations.

For the reverse direction of teaching Chinese as the L2, much computer-assisted language learning (CALL) effort has focused on teaching character writing (Lam, Ki, Law, Chung, Ko, Ho & Pun, 2001; Kuo & Hooper, 2004; Rosell-Aguilar & Kan, 2015), idioms (Luk & Ng, 1998; Wong, Chin, Tan & Liu, 2010; Yang & Xie, 2013), spoken Cantonese (Shī, 2002; Ki, 2006; Wong, 2010), and spoken Mandarin (Chang, 2007). The pedagogical approaches in these studies are not typical of DDL: unlike the present study, none of them was centered on exploring a corpus or making generalizations from corpus examples. One study that is similar to ours in this respect has been reported by Smith, Huang, Kilgarriff and Chen (2008). For six weeks, 25 learners studied Mandarin synonyms and near-synonyms by querying the Chinese Sketch Engine (Kilgarriff, Huang, Rychly, Smith & Tugwell, 2005). In contrast to our work, however, the L1 of most learners was English; the corpus was monolingual; and further, most participants did not complete the post-test, making it difficult to assess the DDL impact. To the best of our knowledge, there is no previous study that evaluates the use of Cantonese-Mandarin parallel corpora.

Advanced, sophisticated learners. Reflecting the general trend in DDL research for language learning (Boulton, 2008a), all cited studies above focused on intermediate to advanced learners. Yet there have been some attempts to bring the benefits of DDL to students at beginner level. In a study by St. John (2001), a beginner student of German was given a parallel German-English corpus and a concordancer as his only tools. Experimental results show that these resources were useful in helping the student tackle various questions on phrases of introduction, modal particles, conjunctions, as well as other grammatical and lexical tasks. Since the student had a linguistics background, however, the approach may not be generally replicable. For teaching English word usage and grammar, Tian (2004) found that DDL leads to greater gains in post-test scores than the conventional textbook approach, for students at both high and low proficiency levels. Focusing on phrasal verbs, Boulton (2008b) showed that lower intermediate students of English can detect their usage patterns in raw data from concordance print-outs. In subsequent experiments on learning linking adverbials and other aspects of English grammar/usage (Boulton, 2008c, 2009, 2010), he found that learners of English at lower levels benefit from DDL materials as much as, if not more than, traditional resources such as dictionary entries and grammar/usage

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

manuals. Pursuing this research direction, our study is concerned with students at the beginner level, with no previous formal instruction in the L2.

Hands-on. Most previous DDL studies on lower level students utilize paper-based materials. Among the few exceptions is the study conducted by Chujo *et al*. (2009), where Japanese university students worked with a Japanese-English parallel newspaper corpus through ParaConc (Barlow, 2004). Our study likewise adopts the hands-on approach by directly exposing students at the beginner level to a parallel corpus through a web-based search interface.

Inductive, self-directed. Chujo *et al*. (2009) followed a four-step approach to teaching English grammar. First, students worked in pairs on worksheets with inductive corpus-based exercises; teachers then confirmed or corrected their hypotheses with explicit explanations. Next, students tested their hypotheses in follow-up exercises; in the final step, teachers again supplied feedback. While our approach is also inductive, it is mostly self-directed, requiring minimal supervision from teachers.

Authenticity. Although authentic material is deemed important for DDL and has been shown to be beneficial for language learning (Larimer & Schleicher, 1999; Erbaggio, Gopalakrishnan, Hobbs & Liu, 2012), simplified corpora are sometimes preferred for students at lower levels in order to accommodate their language proficiency. For example, Chujo, Oghigian, Anthony and Yokota (2013: 67) used the Corpus of Beginner Level English for teaching students at the low remedial level, because "corpora of authentic native speaker language are simply far too difficult". Likewise, Linkit (Shei & Hsieh, 2012) is a CALL system that allows users to search for Mandarin words and phrases with their English translations, drawing material from textbooks and dictionaries rather than naturally occurring texts. Textbook material also forms the basis of Wu and Zhang's (2004) system, which generates language exercises from 35 Mandarin grammar patterns. By contrast, our study uses a parallel corpus with naturally occurring sentences from contemporary media.

3 Methodology

This section begins with a presentation of the parallel corpus used in our study, followed by a description of the experimental setting and the test instruments. We then illustrate the search interface for the corpus and explain how it was used in the DDL activity. Finally, we outline how our experiments were controlled.

3.1 Corpus

Although many Cantonese-Mandarin bilingual dictionaries are available, there is currently only one large-scale parallel corpus of Cantonese and Mandarin Chinese, with more than 75,000 characters (Lee, 2011). In this section, we describe the textual material in this corpus, and summarize its compilation process.

3.1.1 Textual material. Because Cantonese is a predominantly spoken language, it is relatively difficult for its learners to find written samples. Although Cantonese is increasingly used in newspapers and blogs, few supply Mandarin translations. Textbooks may

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Table 1 *Size and source material of Cantonese-Mandarin parallel corpus*

TV Program	Size	# characters
六點半新聞報道	5 episodes × 20 min	19,069 Mandarin
"TVB News at Six-Thirty" (2011)		20,900 Cantonese
溏心風暴之家好月圓	2 episodes × 45 min	17,706 Mandarin
"Moonlight Resonance" (2008)		18,292 Cantonese

contain bilingual sentences, but these are often artificially created, hence not likely to reflect the most colloquial or current usage. Contemporary media are attractive sources because their texts tend to be authentic and natural. From the mid 1990s, TV stations in Hong Kong began to provide Mandarin subtitles on Cantonese TV programs to cater to Mandarin-speaking viewers from mainland China who do not speak Cantonese. The parallel corpus used in this study was constructed from television programs, including news and dramas, broadcast in Hong Kong (see Table 1). Its Cantonese sentences are transcriptions of the TV programs; its Mandarin sentences are the corresponding subtitles.

Table 1 shows the breakdown of the content of the corpus, which contains Cantonese in both formal and colloquial registers. Formal Cantonese often closely resembles Mandarin. As Ramsey (1987: 105) remarks, "almost any Mandarin grammatical pattern can be used in Cantonese and be understood, but such locutions are often not idiomatic… A colloquial Cantonese discourse always has a number of patterns that would sound peculiar in Mandarin." By contrast, idiomatic or colloquial Cantonese tends to diverge from Mandarin to a greater extent in both the lexical and grammatical aspects described in Section 2.1; see Matthews and Yip (2011) for a fuller account. In our corpus, the dramas provide the colloquial register; the news programs contribute mostly to the formal register with the speeches of the anchor and reporters, but also some colloquial register with those of the spontaneous interviewees. With the exception of these spontaneous interviews, all materials consist of pre-planned speech. They are thus largely free of false starts, sentence fragments, repairs, repetitions and other disfluencies, which might be detrimental for pedagogical purposes.

3.1.2 Corpus construction

Sentence alignment. Due to the cost of manual sentence alignment, there is usually a trade-off between large corpus size and automatic alignment (Gao, 2011; Montero Perez *et al*., 2014). The source sentences of this corpus, fortunately, came already aligned: the Mandarin subtitles were aligned to the simultaneously spoken Cantonese.

Word alignment. Word segmentation was automatically performed on the Mandarin sentences (Chang, Galley & Manning, 2008), followed by manual correction. Mandarin words were then manually aligned to their Cantonese counterparts; an example is shown in Figure 1. Word alignments are useful in pedagogy because the correspondence between Cantonese and Mandarin words are not always obvious. For example, the number 'one' can be omitted in front of the classifier + noun construction in Cantonese (e.g., 封信 *fūng seun* '[classifier] letter' in Figure 1); elsewhere, this construction can also imply the definiteness of the noun (letter).

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

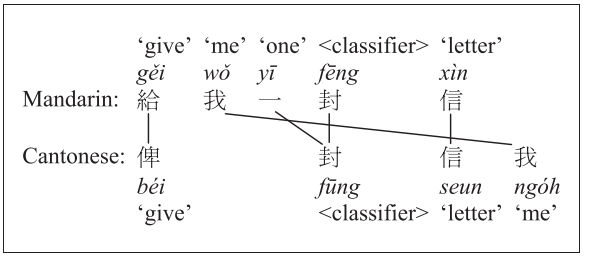


Fig. 1. Example pair of word-aligned Cantonese and Mandarin sentences 'Give me a letter'

Corpus size. The corpus contains 4,135 pairs of such sentences or phrases, with a total of 36,775 characters in Mandarin, and 39,192 in Cantonese. Despite its modest size compared to bilingual corpora of major languages, it exhibits most of the major differences between the two languages (Lee, 2011), and includes some phenomena that are not well represented in textbook sentences, such as code-mixing between English and Cantonese (Lee, 2012).

3.2 Setting and participants

All experiments took place during four semesters between 2012 and 2014 at a thirteen-week course, *Cantonese Communication Skills for Putonghua Speakers*, offered at a university in Hong Kong. The course used a textbook authored by a professor of Chinese at the same university. All participants were Mandarin-speaking undergraduate students, from a variety of majors.[2] Most students had little prior knowledge of Cantonese, and no prior experience with using corpora for language learning.

The outline of the course is provided in Table 2. The design of the DDL and its associated activities was inspired by the language learning model proposed by Nation (2001). In the first stage of this model ('noticing'), the learner must notice the word items to be learned. In the second ('retrieving'), the learner is to recall the items seen in the previous stage. Finally, in the third ('creating, generating'), the learner is to create new contexts or sentences with the items learned. In the seventh week of the course, as the first stage, the pre-test brought to the students' attention their difficulties with any Cantonese word usage; Section 4.2 will provide more detail on the test format. In the ninth week, as the second stage, the CALL session provided repeated encounters of these words in the corpus; Sections 4.3 will describe the DDL activity during the session. Immediately following the session, the post-test measured its short-term effect on the students. In the final week, a delayed post-test measured the long-term effect. Both the post- and delayed post-tests served as the third stage where students applied the newly learned word usage in different contexts.

3.3 Test instruments

Each student completed a pre-test, post-test, and delayed post-test, in the schedule shown in Table 2. Each test contained twelve items, each of which consisted of two Mandarin

[2] Except for Experiment 1, where most participants were business students.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Table 2 *Course outline for Cantonese Communication Skills for Putonghua Speakers*

Week	Lecture topic	DDL and evaluation activities
1	Simple greetings	
2	Getting to know your friends	
3	What time is it?	
4	Happy New Year!	
5	Asking for directions	
6	Let's go dining!	
7	Let's go shopping!	Pre-test
8	Mid-term examination	
9	Let's travel in Hong Kong!	CALL session, post-test
10	Let's go on a picnic!	
11	Seeing the doctor	
12	Student Presentation	
13	Oral examination	Delayed post-test

sentences drawn from our corpus. The two sentences have one word in common, which is underlined. For example, in item (1) in Table 3, the word 沒 *méi* 'have not; not yet' is underlined. The student was asked to translate the underlined Mandarin words into Cantonese *in their specific contexts*. A point was awarded if the student supplied the correct translation in both contexts.

This translation task was designed to test knowledge of Cantonese vocabulary, focusing on cases where the same Mandarin word requires different Cantonese translations depending on its linguistic context – a frequent source of difficulty for beginners who are Mandarin speakers. Continuing with item (1) in Table 3 as our example, when *méi* serves as the verb meaning 'have not' (1-i), it should be translated into 冇 *móuh*; however, when it functions as the adverb meaning 'not yet' (1-ii), it would appear as 未 *meih* in the equivalent Cantonese sentence.

All the underlined words are relatively common, most appearing at least 50 times in the corpus; but they are polysemous, often capable of serving in multiple parts-of-speech – adjectives, adverbs, conjunctions, classifiers, determiners, particles, prepositions, quantifiers, and verbs are all represented in the test. Thus the students were tested on the linguistic functions or senses of these frequently occurring words, rather than their breadth in Cantonese vocabulary.

Table 3 illustrates a few sample items from the pre-test, which was used in all experiments. Similar to item (1) as discussed above, in item (2), 得 *de* remains the same character 得 *dāk* in Cantonese when it marks a complement of manner (2-i), but must be translated as 到 *dou* when it marks a resultative complement (2-ii). In item (3), 雙 *shuāng* can mean 'double', in which case it is identical in Cantonese as 雙 *sēung* (3-i); otherwise, it serves as a classifier, in which case it is equivalent to the Cantonese classifier 對 *deui* (3-ii).

In summary, then, the pre-test assessed whether the student can correctly translate the twelve Mandarin words into Cantonese, each in two different contexts. The post-test and delayed post-test were in the same format, assessing students on the same twelve Mandarin words but using different sentences.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Table 3 *Sample items in pre-test*[3]

		Mandarin sentence	Cantonese translation (model answer)
(1)	(i)	沒辦法	冇
		méi bànfǎ	*móuh*
		have.not solution	have.not
		'That's too bad!'	
	(ii)	還沒睡	未
		hái méi shuì	*meih*
		still not.yet sleep	not.yet
		'[I] haven't gone to bed.'	
(2)	(i)	走得快	得
		zǒu de kuài	*dāk*
		walk PRT quick	PRT
		'[He] walks quickly.'	
	(ii)	笑得流眼淚	到
		xiào de liú yǎnlèi	*dou*
		laugh PRT stream tear	PRT
		'[He] laughs [so much that] tears stream down.'	
(3)	(i)	雙黃白蓮蓉	雙
		shuāng huáng bái liánróng	*sēung*
		double yolk white lotus.seed.paste	double
		'Double-yolk white lotus seed paste [mooncake]'	
	(i)	一雙鞋子	對
		yī shuāng xiézi	*deui*
		one pair shoe	pair
		'a pair of shoes'	

3.4 CALL session

During the one-hour CALL session, students were expected to find Cantonese translations for a list of Mandarin words in context, by consulting the parallel corpus (Section 3) through a web interface (Lee, Hui & Kong, 2013). With this interface, the student can retrieve sentences from the corpus that contain particular Mandarin or Cantonese keywords, then drill into a sentence by viewing its equivalent in the other language; and examine which word(s) in that sentence correspond to the keyword, as derived from word alignments in the corpus. General-purpose corpus visualization platforms, such as ANNIS (Zeldes, Ritz, Lüdeling & Chiarcos, 2009), provide more powerful search capabilities with query languages, but their steep learning curve makes them less suitable for our study in which students were expected to work autonomously. The interface used here offers fewer search options, but requires less training. Other comparable alternatives include ParaConc (Barlow, 2004) and AntPConc (Anthony, 2012), but neither support searching for aligned words.

At the beginning of the session, the teacher briefly demonstrated the corpus search interface. The general workflow runs as follows. After the student has inserted one or

[3] The transliteration, English gloss, and the model Cantonese translation of the underlined Mandarin word were not included in the actual pre-test.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

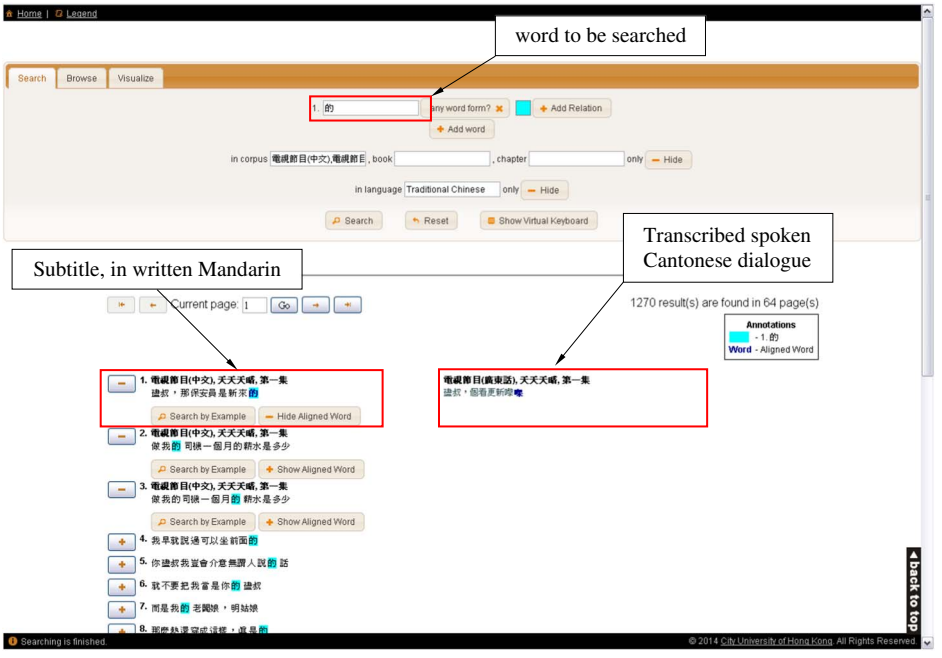


Fig. 2. Web-based search interface used in CALL session

multiple Mandarin keywords, the system retrieves all Mandarin sentences that contain these keywords, and displays them in the keyword-in-context format on the left panel. Figure 2 shows the search results for the keyword 的 *de*. The student can browse these sentences, and select one or more to explore further by pressing the "Show Aligned Word" button. The system then displays the corresponding Cantonese sentence on the right panel, with the aligned word highlighted. For example, in Figure 2, the Cantonese word 㗎 *ga* is highlighted, showing that it corresponds to 的 *de* in the Mandarin sentence. In effect, then, the student exploits the corpus as a contextualized dictionary, a role suggested by Barlow (2000). If the student wants to focus on one particular sense of the keyword, he or she can further restrict search results to only those sentences with a specific aligned word. The same search can be performed using Cantonese keywords.

After the demonstration, the teacher distributed worksheets to the students but gave no further instruction. The worksheet provided a list of Mandarin words, for which the students were asked to find and explain at least two possible Cantonese translations. First, the students searched in the corpus for sentences in which the given Mandarin words appear. Figure 3 illustrates a specific scenario with partial search results for the Mandarin word 了 *le*, which may indicate an aspect or a mood. On the left hand side, four Mandarin sentences are shown with the keyword highlighted. Next, the students read the corresponding Cantonese sentences, and observed the aligned Cantonese words. In Figure 3, the Cantonese equivalents for two of the Mandarin sentences are shown on the right, with the aligned words 喇 *la* and 咗 *jó* bolded. The students would be expected to notice that both of these Cantonese words can possibly correspond to *le*, depending on the linguistic context. After examining as many other sentence pairs as needed to see different

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press



Fig. 3. Partial search results for the Mandarin keyword 了 *le*

contexts, they were expected to identify their major functions or meanings in the worksheet. For example, *jó* denotes the perfect aspect, while *la* is a sentence-final mood particle.

3.5 Control

The ideal experiment for evaluating our pedagogical method would test the same students on the same vocabulary items, and measure any difference in learning outcomes resulting from participation or non-participation in the CALL session. This experimental design is unfortunately infeasible, since a participant cannot learn the same item twice. We thus conducted two experiments with two complementary control methods.

The first experiment used a treatment group which participated in the CALL session, and a control group which did not. We conducted a pre-test to ensure that the initial L2 proficiency levels of both groups were similar. However, differences in personal motivation, effort paid during the CALL session, and outside-class learning activities between the pre-test and the delayed post-test could all potentially confound the results. We report on this experiment in Section 4.1.

The second experiment used treatment items and control items; all students participated in the CALL session. However, the twelve Mandarin words in the pre-test were divided into treatment items and control items, with the latter withheld from the worksheet for the CALL session. We conducted a pre-test to ensure that the two sets of items were of similar difficulty. We then compared their learning outcomes. We report on this experiment in Section 4.2.

4 Experiment Results

In this section, we discuss the results of the two experiments, the first using a control group, the second using control items. We then analyze the students' attitudes towards our pedagogical method, and finally note some limitations of this study.

4.1 Experiment 1

This experiment took place in fall 2012. Of the 55 students who participated, 50 completed all tasks. The 26 students from one class section participated in the CALL session, and thus served as the treatment group; the remaining 24 students from the other section did not participate, and thus served as the control group. A summary of the experimental results is given in Table 4.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Table 4 *Average scores in Experiment 1*

Average score	Treatment group	Control group
Pre-test	66.3%	70.5%
Post-test	86.3%	n/a
Delayed post-test	78.5%	74.5%

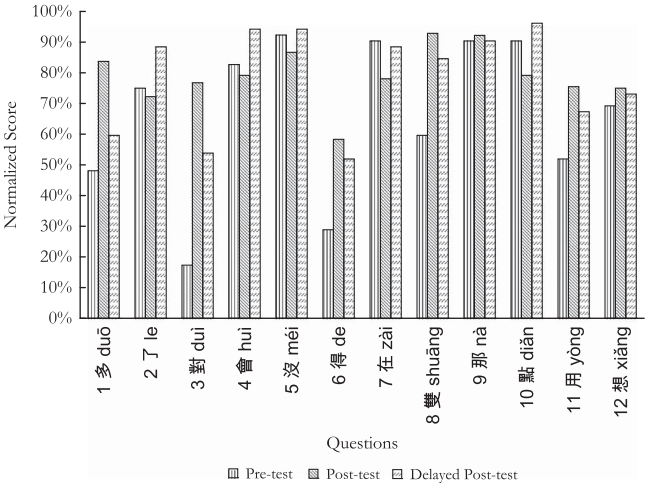


Fig. 4. Scores of treatment group for individual items in Experiment 1

4.1.1 Pre-test. The pre-test was administered to both the treatment and control groups to measure any pre-existing difference in their Cantonese proficiency. As shown in Table 4, the treatment group scored 66.3%, while the control group scored 70.5%. Their difference in performance is not significant ($t = -1.03$, $p = 0.308 > 0.05$, d.f. $= 40$). This shows that the two groups had similar Cantonese proficiency before the experiment.

The average scores of the treatment group for the twelve items in the pre-test are tabulated in Figure 4. The items 得 *de* and 對 *duì* were the most challenging: some students were not aware that a single form in Mandarin could have multiple forms in Cantonese; others recalled one but not both Cantonese translations, or swapped the two.

4.1.2 CALL session, post-test. Only the treatment group participated in the CALL session. Most students seemed to have issued the correct search queries for the Mandarin words on the worksheet, and retrieved the intended Mandarin sentences. At times, they found it difficult to interpret the aligned Cantonese word, often due to differences in grammatical structure or semantic domain (cf. Section 2.1), and sometimes due to unknown words in the rest of the sentence. To be fair to all participants, rather than giving an explanation, the instructor would ask them to look at other examples in the corpus to infer the meaning.

The post-test was given immediately after the CALL session. As shown in Table 4, the performance of the treatment group improved markedly to 86.3%.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Table 5 *Average scores in Experiment 2*

Average score	Treatment items	Control items
Pre-test	34.4%	35.4%
Post-test	85.8%	69.4%
Delayed post-test	74.9%	68.5%

4.1.3 Delayed post-test. In the delayed post-test, the performance of the treatment group decreased slightly, to 78.5%. Despite the decrease, the score was still a statistically significant improvement over their score in the pre-test ($t = -4.85$, $p = 5.46 \times 10^{-5} < 0.0001$, d.f. $= 26$). Compared to the control group, the treatment group also performed better, with an increase of 12.17% to the control group's 4%.[4] The difference between them is significant ($t = 2.27$, $p = 0.0280 < 0.05$, d.f. $= 48$).

Among the twelve items, 對 *duì* and 得 *de*, the lowest-scoring in the pre-test, exhibited the greatest improvement. Conversely, 那 *nà* and 在 *zài*, the highest-scoring in the pre-test, showed no improvement or slightly regressed, possibly due in part to a ceiling effect.

4.2 Experiment 2

This experiment was conducted over three semesters from 2012 to 2014. Of the 101 students who participated, 82 completed all tasks. All figures reported below are aggregates from all semesters. A summary of our experimental results is given in Table 5.

4.2.1 Pre-test. As in Experiment 1, we administered a pre-test to all students. Overall, the students gave correct Cantonese translations in both contexts for 34.9% for the twelve items.

For this experiment, we created personalized worksheets for each student, so that he or she would spend time in the CALL session only on words whose Cantonese translations they had previously failed to master. This means that all Mandarin words for which the student gave correct answers in the pre-test would not appear in his or her worksheet. Among the remainder of the words, we randomly selected half of them as treatment items, to be included in the worksheet; the other half as control items, to be excluded. There were 454 control items and 494 treatment items.

To confirm that the treatment and control items were similarly difficult for the students, we computed their performance on these items in the pre-test. The treatment items scored 34.4%, while the control items scored 35.5%;[5] the difference was statistically insignificant.[6] This suggests that the difficulty levels of the treatment and control items were comparable.

[4] In absolute terms, the increase was 2.92 out of 24 marks for the treatment group and 0.96 out of 24 marks for the control group.

[5] Note that these scores are low because all items for which the student gave correct Cantonese translations are removed from consideration. They are therefore not comparable to their counterparts in Experiment 1.

[6] By chi-square test, $\chi^2 = 0.39$, $p = 0.735 >> 0.05$, d.f. $= 1$.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Table 6 *Sample student responses in worksheets*

Mandarin	Cantonese	Example sentence	Explanation
對 *duì*	啱 *ngāam* 'correct'	唔啱嗰個好似係我嘅喎 'It looks like I was the one who was not correct!'	與"錯"相對 'opposite of "wrong"'
	對 *deui* 'to, towards'	佢係對佢老公忠心啫 'She is just loyal to her husband!'	表示指向某對像 'pointing to some target'
點 *diǎn*	點 *dím* 'o'clock'	聽朝九點一定喺你檯面出現 'You must appear at your counter tomorrow morning at nine!'	時間上的鐘時 'hour'
	啲 *dī* 'bit'	我個肚有啲痛呀要去買藥 'I've got a slight stomach-ache and need to go buy some pills.'	一些、一點 'some, a little bit'

4.2.2 CALL session, post-test. As explained above, each student worked on a personalized worksheet that included his or her treatment items but excluded the control items. In all other aspects, the CALL session was run in the same manner as Experiment 1. Table 6 shows samples from some completed worksheets. Students searched the corpus for two different Cantonese translations of the Mandarin words listed on the worksheet, and gave a brief explanation. In many instances, they were able independently to find the correct Cantonese translations for which they had made an error in the pre-test. For the Mandarin word 對 *duì*, for example, the student successfully identified its two possible Cantonese equivalents, 啱 *ngāam* as the opposite of 'wrong', and 對 *deui* as a preposition.

In the post-test, performance on both the treatment and control items improved substantially. Even for the control items, which were excluded from the worksheet, the score rose to 69.4%. The students likely noticed some of these words in the sentences they browsed, and learned about their usage as a side effect. Meanwhile, the score on the treatment items increased even more sharply, to 85.8%. Hence even with the beneficial side effect, student performance on the treatment items was significantly higher[7] than on the control items, suggesting the effectiveness of the corpus-based approach in the short-term.

4.2.3 Delayed post-test. In the delayed-post test, student performance on the treatment and control items regressed to 74.9% and 68.5%, respectively. However, the score on the treatment items remained significantly better[8] than that of the control items. These results suggest that the use of the parallel corpus also improved students' Cantonese proficiency in the long-term. Figure 5 shows the improvement for individual items from pre-test to

[7] By chi-square test, $\chi^2 = 0.99997$, $p = 1.05 \times 10^{-9} < 5 \times 10^{-9}$, d.f. = 1.

[8] By chi-square test, $\chi^2 = 0.87$, $p = 0.0287 < 0.05$, d.f. = 1.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

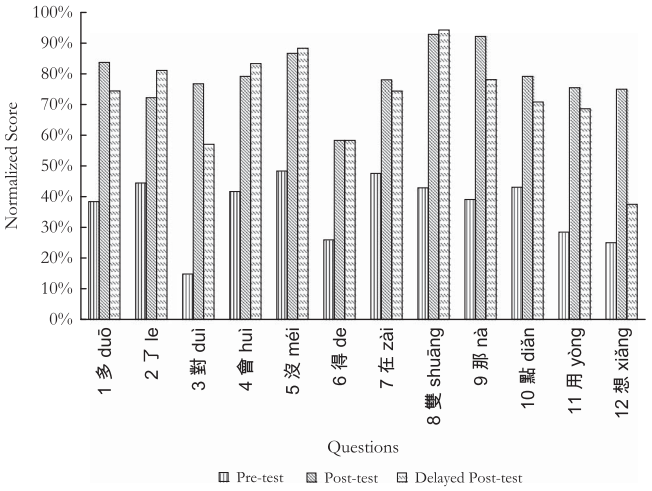


Fig. 5. Scores for individual treatment items in Experiment 2

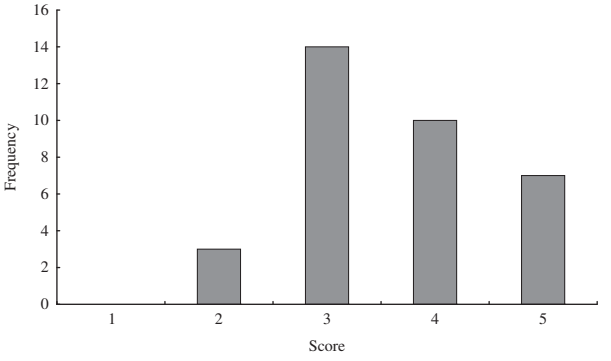


Fig. 6. Score distribution for the answer: "Examples in the corpus are more authentic than those in the textbook" ($\mu = 3.62$, $\sigma = 0.922$, $N = 34$)

delayed post-test. The item 對 *duì* shows the greatest improvement, followed by 想 *xiǎng* and 用 *yòng*.

4.3 Questionnaire results

In order to evaluate the corpus-based approach from the students' perspective, we distributed a questionnaire following the post-test in the last semester of the study. It contained two 5-point Likert scale items (with 1 indicating "strongly disagree", 3 "neutral", and 5 "strongly disagree") and two open-ended questions.

As argued in many previous studies, the use of authentic material is generally considered beneficial for language learning. The first question assessed the authenticity of the sentences in the corpus. The participants appeared divided on this issue. Half of the respondents agreed or strongly agreed that they were more authentic than those in the textbook (Figure 6). The rest of the respondents were mostly neutral on this question. We believe this

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

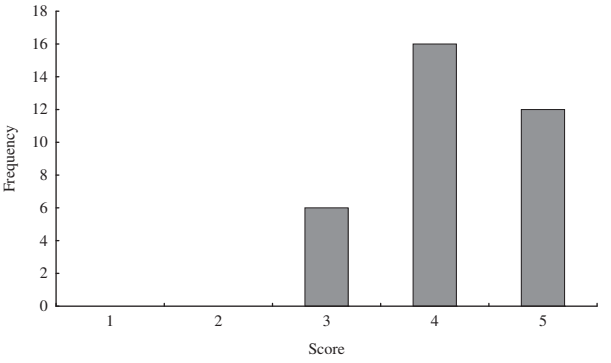


Fig. 7. Score distribution for the answer "The CALL session helps you understand Cantonese usage better" ($\mu = 4.15$, $\sigma = 0.78$, $N = 34$)

reflects the fact that the textbook also includes highly authentic material from colloquial Cantonese, while parts of the corpus also contain more formal or high-register Cantonese.

The second question evaluated the utility of the CALL session on learning Cantonese (Figure 7). Most students gave "4", corroborating our experimental results above indicating that the session significantly improved their Cantonese proficiency. Over a third (35%) gave the maximum "5" for this question.

The third, an open-ended question, asked students what they liked most about the CALL session. 41% mentioned its ability to contrast the two languages, which is a major goal of the parallel corpus. Around 15% praised the authenticity of the sentences, hence suggesting that they succeeded in providing a cultural context in teaching the language, a benefit pointed out by Lange (1999). Others remarked that the sentences were interesting and/or easy to understand.

Finally, when asked how the CALL session may be improved, most students proposed enhancements to the corpus. The most frequent suggestions were to link dictionary entries to, and provide pronunciation guidelines for each word.

4.4 Discussion

Although our experiments suggest that a DDL approach with a Cantonese-Mandarin corpus can be useful for beginner students of Cantonese who are Mandarin speakers, our results should be qualified in several aspects. First, the test instruments measured the impact of DDL in limited scope. They evaluated the participants on an important but rather specific aspect of their language proficiency, namely, the ability to produce appropriate Cantonese translations given a Mandarin word in a particular linguistic context. While this evaluation metric served to directly address the students' DDL task from the CALL session, it left out other aspects of language proficiency, such as grammar and style. Second, beyond language proficiency, DDL may also cultivate other longer-term benefits, such as language awareness and noticing skills (Boulton, 2012). Evaluation on these benefits would present a fuller picture of the impact of the DDL approach.

Despite these limitations, our results contribute to DDL research in several ways. We have shown that beginner students, without the linguistics background enjoyed by the participant in the study by St. John (2001), can also learn effectively in a corpus-based approach. Further, they were able to work autonomously, without the kind of supervision

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

provided in Chujo *et al*. (2009); and to work with a corpus of original material, rather than a simplified one as in Chujo *et al*. (2013). The students' improved performance in post- and delayed post-tests complements results reported by Boulton (2008c, 2009, 2010) and Tian (2004), on a different learning task and in an under-studied L1-L2 pair. It would be interesting to investigate whether the use of monolingual L2 material, an approach taken by Smith *et al*. (2008), can lead to similar performance gains.

It remains an open question whether our findings can be generalized to other L1-L2 pairs that are closely related languages, such as the Romance languages or Slavic languages (cf. Babych, 2015). It may be the case that linguistic similarity alone is not sufficient. To name but one factor, cultural affinity between Cantonese-speaking and Mandarin-speaking societies – as reflected, for example, in the popularity of Cantonese TV programs among Mandarin speakers – may have helped our students in working with the parallel corpus, which contains transcripts from these programs.

5 Conclusions

In this first study of its kind for Cantonese, we have described a corpus-based DDL approach for teaching this language at the beginner level to Mandarin Chinese speakers. This approach centers on a CALL session, where each student actively and independently searches for translation examples in a Cantonese-Mandarin parallel corpus.

We have presented results of two experiments on this approach, involving 131 students and two different types of control. In the first experiment, the treatment group (who participated in the CALL session) outperformed the control group (who did not) in both the post-test and delayed post-test. In the second experiment, both the post-test and delayed post-test showed that the students learned the treatment items more effectively than the control items. Thus, both evaluations suggested that the use of the parallel corpora significantly improved the students' knowledge of Cantonese vocabulary in the long-term. Further, according to the questionnaire results, the students themselves also felt that the corpus helped them understand Cantonese better. Our approach can potentially be replicated not only for other varieties of Chinese, but also for other L1-L2 pairs of sufficient similarity.

For future work, we would like to enrich the Cantonese-Mandarin parallel corpus with pronunciation and syntactic annotations. These additions would enable students to use the corpus-based method to improve their pronunciation, and to search for contrastive grammatical features such as word order.

References

Anthony, L. (2012) *AntPConc*. Tokyo: Waseda University. http://www.antlab.sci.waseda.ac.jp

Babych, S. (2015) Textual cohesion patterns for developing reading skills: A corpus-based multilingual learning environment. In Leńko-Szymańska, A. and Boulton, A. (eds.), *Multiple affordances of language corpora for data-driven learning*. 155–175. Amsterdam: John Benjamins.

Barlow, M. (2000) Parallel texts in language teaching. In Botley, S. P., McEnery, M. A. and Wilson A. (eds.), *Multilingual corpora in teaching and research*. Amsterdam/Atlanta: Rodopi, 107–115.

Barlow, M. (2004) *ParaConc*. http://www.athel.com/para.html

Boulton, A. (2008a) But where's the proof? The need for empirical evidence for data-driven learning. In Edwardes, M. (ed.), *Proceedings of BAAL annual conference 2007*. London: Scitsiugnil Press, 13–16.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Boulton, A. (2008b) Looking (for) empirical evidence of data-driven learning at lower levels. In Lewandowska-Tomaszczyk, B. (ed.), *Corpus linguistics, computer tools, and applications: State of the art*. Frankfurt: Peter Lang, 581–598.

Boulton, A. (2008c) DDL: Reaching the parts other teaching can't reach? *Proceedings of the teaching and language corpora conference*. Lisbon: Associação de Estudos e de Investigação Cientifíca do ISLA-Lisboa, 38–44.

Boulton, A. (2009) Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, **21**(1): 37–54.

Boulton, A. (2010) Data-driven learning: Taking the computer out of the equation. *Language learning*, **60**(3): 534–572.

Boulton, A. (2011) Data-driven learning: The perpetual enigma. In Goźdź-Roszkowski, S. (ed.), *Explorations across languages and corpora*. Frankfurt: Peter Lang, 563–580.

Boulton, A. (2012) Language awareness and medium-term benefits of corpus consultation. In Gimeno Sanz, A. (ed.), *New trends in CALL – working together*. London: Macmillan, 39–46. https://hal.archives-ouvertes.fr/hal-00502606v2/document

Chan, T.-P. and Liou, H.-C. (2005) Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning*, **18**(3): 231–250.

Chang, L. L. (2007) The effects of using CALL on advanced Chinese foreign language learners. *CALICO Journal*, **24**(2): 331–353.

Chang, P.-C., Galley, M. and Manning, C. D. (2008) Optimizing Chinese word segmentation for machine translation performance. In Callison-Burch, C., Koehn, P., Monz, C. and Fordyce, C. S. (eds.), *Proceedings of the 3rd workshop on statistical machine translation*. Stroudsbury: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1626430

Chujo, K., Anthony, L. and Oghigian, K. (2009) DDL for the EFL classroom: Effective uses of a Japanese-English parallel corpus and the development of a learner-friendly, online parallel concordancer. In Mahlberg, M., González-Díaz, V. and Smith, C. (eds.), *Proceedings of the corpus linguistics conference (CL 2009)*. University of Liverpool. http://ucrel.lancs.ac.uk/publications/cl2009/48_FullPaper.doc

Chujo, K., Oghigian, K., Anthony, L. and Yokota, K. (2013) Teaching remedial grammar through data-driven learning using AntPConc. *Taiwan International ESP Journal*, **5**(2): 65–90.

Erbaggio, P., Gopalakrishnan, S., Hobbs, S. and Liu, H. (2012) Enhancing student engagement through online authentic materials. *The International Association for Language Learning Technology Journal*, **42**(2): 27–51.

Gao, Z.-M. (2011) Exploring the effects and use of a Chinese-English parallel concordancer. *Computer Assisted Language Learning*, **24**(3): 255–275.

Geist, M. and Hahn, A. (2012) Using a corpus for written production: A classroom study. In Thomas, J. E. and Boulton, A. (eds.), *Input, process and product: Developments in teaching and language corpora*. Brno: Masaryk University Press, 123–135.

Hannas, W. C. (1997) *Asia's orthographic dilemma*. Honolulu: University of Hawaii Press.

Huang, H.-T. and Liou, H.-C. (2007) Vocabulary learning in an automated graded reading program. *Language Learning & Technology*, **11**(3): 61–82.

Johns, T. F. (1991) Should you be persuaded: Two samples of data-driven learning materials. In Johns, T. F. and King, P. (eds.), *Classroom concordancing. English Language Research Journal*. Birmingham: Birmingham University, **4**: 1–13.

Johns, T. F. (1994) From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In Odlin, T. (ed.), *Perspectives on pedagogical grammar*. Cambridge: Cambridge University Press, 293–313.

Johns, T. F. (1997) Contexts: The background, development and trialling of a concordance-based CALL program. In Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds.), *Teaching and language corpora*. London: Longman, 100–115.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Ki, W. W. (2006) Computer-assisted perceptual learning of Cantonese tones. *The 14th international conference on computers in education*. Peking Normal University, 30/11/06.

Kilgarriff, A., Huang, C., Rychly, P., Smith, S. and Tugwell, D. (2005) Chinese word sketches. In Ooi, B. Y. V., Pakir, A., Talib, I. B. S., Tan, L., Tan, K. W. P. and Tan, Y. Y. (eds.), *Words in Asian cultural context: Proceedings of the 4th ASIALEX conference*. National University of Singapore, 1–3/06/05.

Kuo, M.-L. A. and Hooper, S. (2004) The effect of visual and verbal coding mnemonics on learning Chinese characters in computer-based instruction. *Educational Technology Research and Development*, **52**(3): 23–34.

Lam, H. C., Ki, W. W., Law, N., Chung, A. L. S., Ko, P. Y., Ho, A. H. S. and Pun, S. W. (2001) Designing CALL for learning Chinese characters. *Journal of Computer Assisted Learning*, **17**(1): 115–128.

Lange, D. L. (1999) Planning for using the new national culture standards. In Phillips, J. and Terry, R. M. (eds.), *Foreign language standards: Linking research, theories, and practices*. Lincolnwood, IL: National Textbook Company, 57–120.

Larimer, R. E. and Schleicher, L. (eds.) (1999) *New ways in using authentic materials in the classroom*. Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.

Lee, J. (2011) Toward a parallel corpus of spoken Cantonese and written Chinese. In Wang, H. and Yarowsky, D. (eds.), *Proceedings of the 5th international joint conference on natural language processing*. Chiang Mai: Asian Federation of Natural Language Processing. https://aclweb.org/anthology/I/I11/I11-1174.pdf

Lee, J. (2012) Corpus-based analysis of mixed code in Hong Kong speech. In Xiong, D., Castelli, E., Dong M. and Yen P. T. N. (eds.), *Proceedings of 2012 international conference on Asian language processing*. Hanoi: IEEE.

Lee, J., Hui, C. Y. and Kong, Y. H. (2013) Treebanking for data-driven research in the classroom. In Derzhanski, I. and Radev D. (eds.), *Proceedings of the 4th workshop on teaching natural language processing*. Stroudsburg: Association for Computational Linguistics. https://www.aclweb.org/anthology/W/W13/W13-3409.pdf

Li, D. C. S., Wong, C. S. P., Leung, W. M. and Wong, S. T. S. (2016) Facilitation of transference: The case of monosyllabic salience in Hong Kong Cantonese. *Linguistics*, **54**(1).

Luk, R. W. P. and Ng, A. B. Y. (1998) Computer-assisted learning of Chinese idioms. *Journal of Computer Assisted Learning*, **14**(1): 2–18.

Mair, V. H. (1991) What is a Chinese "dialect/topolect"? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers*, **29**: 1–31.

Matthews, S. and Yip, V. (2011) *Cantonese: A comprehensive grammar*. New York: Routledge.

Montero Perez, M., Paulussen, H., Macken, L. and Desmet, P. (2014) From input to output: the potential of parallel corpora for CALL. *Language Resources and Evaluation*, **48**(1): 165–189.

Nation, I. S. P. (2001) *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nerbonne, J. (2000) Parallel texts in computer-assisted language learning. In Veronis, J. (ed.), *Parallel text processing*. Dordrecht and Boston: Kluwer, 354–369.

Ōuyáng, J. (1993) *Pǔtōnghuà Guǎngzhōuhuà de bǐjiào yǔ xuéxí (The comparison and learning of Mandarin and Cantonese)*. Peking: China Social Science Press.

Poole, R. (2012) Concordance-based glosses for academic vocabulary acquisition. *CALICO Journal*, **29**(4): 679–693.

Ramsey, S. R. (1987) *The languages of China*. Princeton: Princeton University Press.

Rosell-Aguilar, F. and Kan, Q. (2015) Design and user evaluation of a mobile application to teach Chinese characters. *JALT CALL journal*, **11**(1): 19–40.

St. John, E. (2001) A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning & Technology*, **5**(3): 185–203.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press

Shei, C. and Hsieh, H.-P. (2012) Linkit: a CALL system for learning Chinese characters, words, and phrases. *Computer Assisted Language Learning*, **25**(4): 319–338.

Shī, Z. (2002) *Guǎngzhōu yīn Běijīng yīn Duìyìng Shǒucè (A handbook on the correspondence between Cantonese pronunciation and Pekinese pronunciation)*. Canton: Jinan University Press.

Smith, S., Huang, C.-R., Kilgarriff, A. and Chen, M.-R. (2008) A corpus query tool for SLA: Learning Mandarin with the help of Sketch Engine. In Lewandowska-Tomaszczyk, B. (ed.), *Corpus linguistics, computer tools, and applications – state of the art*. Frankfurt: Peter Lang, 673–686.

Tadmor, U., Haspelmath, M. and Taylor, B. (2010) Borrowability and the notion of basic vocabulary. *Diachronica*, **27**(2): 226–246.

Tian, S. (2004) Data-driven learning: Do learning tasks and proficiency make a difference? In *Proceedings of the 9th conference of pan-Pacific association of applied linguistics*. http://www.paaljapan.org/resources/proceedings/PAAL9/pdf/TianShiaup.pdf

Tono, Y., Satake, Y. and Miura, A. (2014) The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, **26**(2): 147–162.

Wang, L. (2001) Exploring parallel concordancing in English and Chinese. *Language Learning & Technology*, **5**(3): 174–184.

Wong, L.-H., Chin, C.-K., Tan, C.-L. and Liu, M. (2010) Students' personal and social meaning making in a Chinese idiom mobile learning environment. *Educational Technology & Society*, **13**(4): 15–26.

Wong, T.-S. (2010) A pilot study on the outcome of teaching phonological correspondence in Cantonese class for Mandarin speakers. *The 2010 Annual research forum of the linguistic society of Hong Kong (LSHK-ARF 2010)*. The Chinese University of Hong Kong, 01/12/10.

Wu, Y. and Zhang, J. (2004) A Chinese language expert system using Bayesian learning. In Callaos, N., Lesso, W. and Sanchez, B. (eds.), *Proceedings of the 8th world multiconference on systemics, cybernetics and informatics*, Florida. http://facultyweb.cs.wwu.edu/~zhangj/home/papers/sci04-nlp.pdf

Yang, C. and Xie, Y. (2013) Learning Chinese idioms through iPads. *Language Learning & Technology*, **17**(2): 12–23.

Zeldes, A., Ritz, J., Lüdeling, A. and Chiarcos, C. (2009) ANNIS: A search tool for multi-layer annotated corpora. In Mahlberg, M., González-Diaz, V. and Smith C. (eds.), *Proceedings of the corpus linguistics conference (CL2009)*. University of Liverpool, 20–23/07/09. http://ucrel.lancs.ac.uk/publications/cl2009/358_FullPaper.doc

Zeng, Z. (1993) *Colloquial Cantonese and Putonghua equivalents (4th edn.)*. (S. K. Lai, trans.). Hong Kong: Joint Publishing (Hong Kong) Company Limited.

https://doi.org/10.1017/S0958344015000257 Published online by Cambridge University Press