ReCALL (2023), 35: 1, 19–39
doi:10.1017/S0958344022000180

RESEARCH ARTICLE

Teachers' perceived corpus literacy and their intention to integrate corpora into classroom teaching: A survey study

Qing Ma
The Education University of Hong Kong, Hong Kong SAR, China (maqing@eduhk.hk)

Ming Ming Chiu
The Education University of Hong Kong, Hong Kong SAR, China (mingmingchiu@gmail.com)

Shanru Lin
The Education University of Hong Kong, Hong Kong SAR, China (lllam32316@gmail.com)

Norman B. Mendoza
The Education University of Hong Kong, Hong Kong SAR, China (normanmendoza0421@gmail.com)

Abstract

Given the importance of corpus linguistics in language learning, there have been calls for the integration of corpus training into teacher education programmes. However, the question of what knowledge and skills the training should target remains unclear. Hence, we advance our understanding of measures and outcomes of teacher corpus training by proposing and testing a five-component theoretical framework for measuring teachers' perceived corpus literacy (CL) and its subskills: understanding, search, analysis, and the advantages and limitations of corpora. Also, we hypothesised that teacher CL is linked to their intention to use corpora in classroom teaching. Specifically, 183 teachers and student teachers received corpus training to develop their CL and then completed a survey to measure their CL and intention to use corpora in teaching in Likert-scale items together with open-ended questions. Confirmatory factor analysis indicated that a hierarchical factor structure for CL using the aforementioned five subfactors best fitted the data. Moreover, structural equation modelling indicated that CL is positively linked to the participants' intention to integrate corpora into classroom teaching. While all five subskills are important for teachers, greater effort should be made to develop their corpus search and analysis skills, which can be viewed as the "bread and butter" of corpus training.

Keywords: corpus literacy; teacher intention; teacher education; classroom teaching

1. Introduction

Corpus linguistics is perceived to have revolutionised language learning, since learners are encouraged to become researchers who can deduce language rules from fresh, rich, and authentic corpus data (Boulton & Cobb, 2017; Johns, 1991; Sinclair, 1991). Corpora have been used extensively in nearly all branches of language studies, including lexicography, grammar, semantics,

Cite this article: Ma, Q., Chiu, M.M., Lin, S. & Mendoza, N.B. (2023). Teachers' perceived corpus literacy and their intention to integrate corpora into classroom teaching: A survey study. *ReCALL* 35(1): 19–39. https://doi.org/10.1017/S0958344022000180

pragmatics, language variation/change, contrastive and translation studies, and discourse analysis (McEnery & Xiao, 2011). Accordingly, corpus linguistics has greatly advanced our understanding of language analysis, shifting from intuition- to corpus-based inductive observations of language accounts. However, its influence on language teaching remains marginal (Boulton, 2017; Callies, 2019; Ma, Tang & Lin, 2021).

Despite continuous calls for the integration of corpus training into teacher education programmes (Breyer, 2009; Callies, 2019; Chen, Flowerdew & Anthony, 2019; Farr, 2008; Leńko-Szymańska, 2017), a key issue remains to be solved – what skills and knowledge teachers should develop in the training. This study argues that corpus literacy (CL) needs to be developed before teachers can develop the necessary pedagogical skills to integrate corpora into their teaching. Building on the work of Mukherjee (2006) and Callies (2019), we proposed a five-component CL framework and collected survey data from 183 teachers and student teachers to test whether improving CL would increase teacher intention to use corpora in their teaching. No previously published study has conducted a comprehensive survey to verify the specific skills and knowledge involved in the CL construct. In addition, we explored teacher-perceived advantages and limitations of integrating corpora into classroom teaching.

A structural equation modelling (SEM) approach was adopted to analyse the collected Likert-scale survey data. Further, the teacher-perceived benefits and limitations of corpora collected from open-ended questions were analysed qualitatively. Our study adds empirical evidence to support CL as a key theoretical construct, which contributes to our theoretical understanding of corpus research on teacher training. The findings of this research can also provide practical guidance for various stakeholders (corpus linguists, corpus website developers, and teacher educators) to inform them how to provide effective corpus teacher training.

2. Literature

2.1 Corpus linguistics in language education

Scholars began using corpus linguistics as a research tool in the 1980s, and educators subsequently discovered its practical applications for language learning and teaching (Johns, 1991; Sinclair, 1991). Leech (1997) outlined two types of corpus applications for language teaching: (1) using corpora during classroom teaching (*direct use*) and (2) aiding teaching and teacher development through the development of corpus-based references/dictionaries, teaching materials, or testing materials (*indirect use*). Several decades after the calls to use corpora for teaching, studies have documented evidence of both indirect use, primarily involving reference publishing (Leńko-Szymańska & Boulton, 2015; McEnery & Xiao, 2011), and direct use of integrating corpora into language teaching (Boulton & Cobb, 2017; Boulton & Vyatkina, 2021; Lee, Warschauer & Lee, 2017; Pérez-Paredes, 2019). However, there is insufficient evidence of direct use regarding the professional development of corpus training for in- and pre-service teachers (Callies, 2019; Latif, 2021; Leńko-Szymańska, 2017; Naismith, 2017; Pérez-Paredes, 2022; Schmidt, 2022), particularly those at primary or secondary schools.

While corpus linguistics and corpora use have energised language research in recent decades, few teachers have integrated them into their classroom teaching, partially due to the absence of in- and pre-service teacher training (Boulton, 2017; Breyer, 2009; Callies, 2019; Leńko-Szymańska, 2017). Teachers and students also report difficulties in using corpus technology (Fitzgerald, 2018; Poole, 2022). Students with limited language proficiency often lack sufficient metalinguistic knowledge to formulate corpus search queries (Chang, 2014; Yeh, Liou & Li, 2007; Yoon & Hirvela, 2004). In addition, messy or overwhelming concordance results may have hindered students from locating language features (Chen, 2011; Rodgers, Chambers & Le Baron-Earle, 2011). Similarly, teachers often reported technical problems regarding how to use corpora, such as searching for, analysing, or interpreting concordance lines (Breyer, 2009; Farr, 2008;

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Leńko-Szymańska, 2014; Zareva, 2017). Furthermore, many teacher participants lacked the requisite knowledge or skills to use corpora independently, even after corpus training (O'Keeffe & Farr, 2003; Tribble, 2015).

2.2 Teacher knowledge and skills for integrating corpora into teaching

Despite the small number of studies on corpus teacher training (Breyer, 2009; Farr, 2008; Heather & Helt, 2012; Leńko-Szymańska, 2017; Zareva, 2017) that demonstrated varying levels of success, a key issue remains for corpus educators: to determine the skills and knowledge that should be focused on in training programmes (Callies, 2016; Heather & Helt, 2012; Mukherjee, 2006). Here, the concept of CL, which was raised by Mukherjee (2006), could be used as a starting point to understand the essential knowledge and skills desired in corpus training. Mukherjee's CL included four components: (1) a basic understanding of what a corpus is, (2) what one can (and cannot) do with a corpus, (3) how concordances can be analysed, and (4) how one may (or may not) extrapolate general trends in language use from corpus data. Callies (2016) further proposed a *searching corpora* component – how to search corpora using corpus resources and tools. To develop CL, one needs to work with a corpus and use a concordancer, which allows the user to enter a keyword/phrase and obtain many search results formulated as concordance lines. Heather and Helt (2012) formally defined CL as the ability to use corpus resources/technology to carry out language analysis.

Different researchers (e.g. Heather & Helt, 2012; Leńko-Szymańska, 2017; Zareva, 2017) have proposed certain knowledge and skills that teachers should develop in corpus training. Although the knowledge and skills can overlap and are different (to a certain extent), a closer examination reveals that they all fall into two broad areas: (1) CL that is similar to that proposed by Mukherjee (2006) and (2) some pedagogy-related skills that aim to integrate corpora into language classrooms (Breyer, 2009; Leńko-Szymańska, 2017; Lin & Lee, 2015). In Ma *et al.* (2021), the latter is formally referred to as corpus-based language pedagogy (CBLP), meaning "the ability to integrate corpus linguistics technology into classroom language pedagogy to facilitate language teaching" (p. 2). In our view, CL relates primarily to content knowledge, whereas CBLP is more concerned with pedagogical content knowledge, based on Shulman's (1987) work.

Such a differentiation of the two dimensions of teacher knowledge is supported by recent empirical work that explored corpus teacher training. For example, the study conducted by Leńko-Szymańska (2017) included both corpus linguistic and pedagogical skills for pre-service teachers. Further, in the study by Çalışkan and Kuru Gönen (2018), teachers initially experienced using corpus as a learning tool and then as a teaching tool. The pre-service teachers in the study by Ebrahimi and Faghih (2017) were first instructed on the use of corpora and then on how to design pedagogical applications of corpora. Such sequence of the training content suggests that CL precedes CBLP, meaning the pedagogical applications of corpora in classroom teaching depend on the CL acquired by the trainees. Through being aware of the complex and intriguing relationship between CL and CBLP, this research focused on CL and aimed to identify the key components of CL. This could provide a sound foundation for teachers to develop their pedagogical skills further to apply corpora in teaching.

2.3 Empirical studies measuring teachers' CL

With regard to the measurement of teacher CL, the predominant approach is to use qualitative methods, including interviews, open-ended questions, reflections, and lesson analysis (e.g. Breyer, 2009; Heather & Helt, 2012; Latif, 2021; Leńko-Szymańska, 2017; Zareva, 2017). A few studies have also used survey-based approaches, including both closed- and open-ended questions (Callies, 2019; Farr, 2008; Leńko-Szymańska, 2014).

The majority of the aforementioned studies focused on measuring the attitudes and perceptions of teachers and student teachers towards the use of corpora in learning and teaching. For example, the experiences of corpus training received by participants have been examined

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

(Breyer, 2009; Ebrahimi & Faghih, 2017; Leńko-Szymańska, 2014; Zareva, 2017). Participant attitudes towards using corpus resources and tools for learning and teaching purposes have also been investigated extensively (Çalışkan & Kuru Gönen, 2018; Farr, 2008; Latif, 2021; Mukherjee, 2006; Naismith, 2017).

Recent research has indicated that corpus skills are multifaceted, including understanding and working with concordance lines (e.g. Charles, 2018; Pérez-Paredes & Mark, 2021) and building corpora (e.g. Charles, 2012, 2014; Charles & Hadley, 2022). Corpora are often built for classroom-specific contexts of language teaching or learning, especially for ESP/EAP instruction at tertiary level where corpora can be tailored to help university lecturers identify student-specific language use, such as analysing ESP/EAP writing tasks (Ackerley, 2021; Chang, 2014; Charles, 2012). Unfortunately, the majority of teachers and student teachers working in school settings have little corpus knowledge, and we believe it is essential to help them build a basic CL, as outlined by Mukherjee (2006) and Callies (2016). Building corpora is considered an advanced corpus skill that can be focused on at a later stage, when school teachers have gained familiarity and developed confidence with their use.

Since the majority of our participants are from school settings, we focus on and measure all the basic CL components outlined by Mukherjee (2006) and Callies (2016): *understanding*, *advantages* and *disadvantages*, *analysis*, and *search skills* of corpora. Among these components, understanding is a relatively easy component to measure, where relevant questions (open-ended or Likert scale) can be designed to measure the level of understanding of corpora before and after training (Leńko-Szymańska, 2014; Mukherjee, 2006; Özbay & Kayaoğlu, 2015). Occasionally, the perceived advantages and disadvantages of corpora have also been explored (Callies, 2019; Lin & Lee, 2015; Mukherjee, 2006; Zareva, 2017). Similarly, participants' perceived confidence (or challenges) in corpus analysis and their interpretation skills of corpus data have also been examined (Breyer, 2009; Leńko-Szymańska, 2014; Zareva, 2017).

2.4 Issues to be addressed and research questions

Past studies (often with small samples) have identified one or two CL subskills but have not comprehensively explicated a full system of critical CL skills. Leńko-Szymańska (2014) measured the understanding component of CL by asking 13 trainee teachers to define two corpus-related terms, such as concordance and concordancer, via open questions. Similarly, Callies (2019) asked 26 teachers to complete a survey with Likert-type scales to investigate three aspects of corpus use in the classroom: their familiarity with corpus applications (entailing understanding), their actual use of corpora in their language teaching, and the advantages of corpus use. Using qualitative data, Heather and Helt (2012) investigated six student teachers about four CL components: understanding, advantages, limitations, and analysis. To examine all CL subskills thoroughly, this larger study surveys many participants regarding all five components of CL: understanding, search, analysis, and the advantages and limitations.

In addition to the above observations, teacher training studies often include training in using corpora for both learning and teaching purposes. Hence, theoretically, participants are expected to use (or at least have the intention to use) corpora in their teaching after the training. Empirically, some small-scale studies (working with a small number of participants and using primarily qualitative data) have revealed that teachers or student teachers were less motivated to apply corpora in classroom after the corpus training (Ebrahimi & Faghih, 2017; Latif, 2021; Lin & Lee, 2015; Zareva, 2017). Moreover, similar to measuring CL, the intention to use corpora in classroom teaching has rarely been investigated systematically. In this study, we work with a large sample of teachers/student teachers and investigate their intentions to adopt corpora in classroom teaching after corpus training.

Finally, Heather and Helt's (2012) findings revealed that participants who had developed good CL were more likely to identify the limitations of corpus linguistics and propose possible solutions.

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

However, relatively few studies have evaluated the limitations of corpora among teachers, partly because an examination of corpus limitations is perceived to be most challenging during corpus teacher training (Zareva, 2017).

To address the aforementioned issues, we adopted a Likert-type scale survey to measure teachers' self-reported CL after receiving corpus training. Our large survey study of 183 participants aims to verify empirically the five subskills of CL proposed by Mukherjee (2006) and Callies (2019) to gather evidence regarding the structure of the CL as an important theoretical construct, thus advancing our understanding regarding corpus research on teacher education. Our study may also shed light on how to provide effective corpus training for teachers. The following research questions guide this study:

1. What are the key subskills involved in developing CL?
2. Do participants with greater CL display a stronger intention to integrate corpora into classroom teaching?
3. What do participants perceive as the benefits and limitations of exploiting corpora in classroom teaching?

3. Methods

3.1 Context and participants

Using convenience sampling, corpus training was provided for 410 teachers and student teachers who were enrolled in courses offered by a university in Hong Kong from 2018 to 2020. Since this was an exploratory study, no control was imposed on the sample distribution. At the end of the corpus-based training, the participants were invited to complete an online survey to measure their perceived CL and their intention to integrate corpora into classroom teaching. Since the survey was voluntary and conducted outside the training, only 183 participants completed the survey, representing a response rate of 44%. Of these participants, most were of Chinese origin (either from Hong Kong or mainland China), while approximately 5% were international students from other countries (such as Japan, the Philippines, the United States, or Canada). Approximately half ($n = 94$) were English teachers, while the other half were a mixture of undergraduate and postgraduate student teachers ($n = 89$). All student teachers were pursuing programmes related to English language education and were expecting to become English teachers in primary or secondary schools after graduation. A large proportion of the in-service teachers (59%) worked in secondary schools, some (30%) worked in primary schools, and a small proportion (11%) worked in tertiary institutions. Accordingly, the majority of the teachers and student teachers had similar needs in terms of teaching English to either secondary or primary schools, justifying the decision to provide them with similar corpus training. Teachers with doctoral degrees often were more familiar with corpus technology, compared to other teachers (Pérez-Paredes, Ordoñana Guillamón & Aguado Jiménez, 2018). Although we did not collect data on participants' educational background, our personal knowledge of them suggests that less than 3% of the participants held a doctoral degree. Our pre-survey also showed that, prior to attending our corpus training, 50% had never heard of a corpus, and only 11% had more than one year of experience with corpora. Hence, most of our participants had limited CL knowledge or skills. Please see Table 1 for a list of the participants' demographic information.

3.2 Procedure

These participants received corpus training following one of two procedures: integrated into one course on vocabulary learning and teaching for student teachers or in stand-alone training programmes offered to teachers. All the training was conducted by the first researcher of this article, who is a CALL researcher specialising in corpus technology and has been working on

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Table 1. Information on participants who completed the survey ($N = 183$)

Variable	Specification	Frequency	Percentage
Teacher status	In-service	94	51%
	Pre-service	89	49%
School level *(applicable to teachers only)*	Primary	28	30%
	Secondary	55	59%
	Tertiary	11	11%
Study level *(applicable to student teachers only)*	Undergraduate	49	55%
	Postgraduate	40	45%
Origin	Hong Kong	80	44%
	Mainland China	86	47%
	Others	17	9%

practical corpora applications with pre- and in-service English teachers during the past few years. In both cases, the participants underwent similar training in terms of length (approximately 4 weeks) and amount of learning content. Since the majority of our participants were already teaching (or would become teachers) at primary or secondary schools, the corpus training will focus on corpus resources and skills suitable for these two settings. Table 2 provides details of the training procedure.

Although the training covered both CL and pedagogical skills, we focus on CL in this study. With regard to the CL training, workshops and independent online learning tasks were provided for the participants to help them develop their understanding of basic concepts relating to corpus and concordance lines: how to search free online corpora (e.g. Lextutor: https://www.lextutor.ca/conc/; Word and Phrase: https://www.wordandphrase.info/; and COCA: http://corpus.byu.edu/coca/). First, corpus searches were demonstrated, and then participants were provided with hands-on experience in performing various corpus searches. Taking COCA as an example, all the essential search functions (including lists, wildcards, parts of speech, collocates, and comparisons) were introduced, and the participants were then given opportunities to practise using the search functions. Along with learning how to use various search functions for different corpora, the participants were guided in analysing and summarising the associated language patterns. Finally, the participants were provided with opportunities to reflect on the advantages and limitations of corpora usage in relation to teaching and learning.

A survey was administered to the participants at the end of the training that contained both Likert-scale and open-ended questions. The data collected from the 183 participants who responded to our survey formed the source for data analysis.

3.3 Constructing CL and developing measurement items

Mukherjee (2006) suggested that CL comprised four components: (a) understanding of corpora, (b) knowing what corpora use can and cannot achieve, (c) knowing how to analyse corpus data, and (d) knowing how to draw conclusions about language use trends. For practical reasons, we divided the second component into two: the limitations and advantages of using corpora. Being able to draw conclusions about language use upon observations of corpus data is a salient advantage of corpora use. Therefore, we decided to integrate this component with the advantages component. When exploring any corpus data, the key corpus tool is the concordancer. However, as indicated by Naismith (2017), “use of concordancers also typically requires some expertise and training, and the results [of concordance lines] may be challenging to interpret” (p. 275). We

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Table 2. Training procedure

Week	Topic	Workshop (2 hours)	Online learning (Schoology)
1	Introduction to corpora	Definition of corpus Demonstration of corpus search in Lextutor and Word and Phrase Participant hands-on corpus searches Brainstorming ideas of applying corpora in learning and teaching	Hands-on corpus searches on Lextutor and Word and Phrase Searching for collocations Summarising language patterns Forum discussion
2	Enhancing student vocabulary learning through corpora	Introduction to COCA Demonstration of COCA search functions (List, Wildcard, Part of Speech, Synonym, Collocates, KWIC, Compare, Chart)	Video watching (COCA) Quiz questions (COCA searches) Forum discussion
3	Integrating corpora into classroom teaching	Design principles of corpus-based learning activities Analysis of corpus-based lesson materials Discussion of advantages and limitations of corpora	Participant design of corpus-based lessons on a self-chosen topic Forum discussion
4	Designing corpus-based lessons	Nil	Uploading the lesson for sharing Commenting on others' lessons and revising their own lesson

believe that teachers and students should be provided with adequate search skills to enable them to perform corpus searches and to observe and analyse language use. Callies (2016) proposed an additional component (*searching corpora*), which we further modified and termed *corpora search skills*. After proposing the five subskills to be included in CL, the next step was to develop relevant scale items targeting each of the key CL components. This resulted in a total of 16 items across five components of CL (see Table 3 for detailed items).

Participants could answer on a 6-point Likert scale ranging from "strongly disagree" to "strongly agree" for all the items. In addition, the following two open-ended questions were included in the survey to collect teacher-perceived benefits and limitations of corpora: (1) In your view, what are the main advantages of corpus knowledge for your language learning and teaching? and (2) What are the limitations of using corpus resources for language learning and teaching?

We initially piloted these items with 40 student teachers and then revised the wording of a few items after receiving student feedback, and all 16 items were retained (see Table 3 for the finalised items).

Although the development of CL may not result in teachers immediately integrating corpora into the classroom, we postulated that the development of CL could have a positive influence on teachers' intention to integrate corpora into classroom teaching. Following the concept of "Intention to Use Technology" in the technology acceptance model (TAM) (Davis, 1989), intention to integrate corpora into classroom teaching was operationalised as a holistic propensity to integrate them into classroom teaching. Accordingly, our instrument also had to consider developing scale items with respect to the intention to integrate corpora into classroom teaching (ITICT). For this purpose, we added three items to form this component (see Table 4 for details).

In essence, the survey instruments were developed to measure two areas of corpus training: (1) the five components to be included in CL development and (2) whether CL can result in ITICT.

3.4 Data analysis

3.4.1 Item analysis

The corrected item-total correlation method was used for item analysis. The correlation between an item and its subscale was considered weak if the correlation was below 0.40, indicating that the

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Table 3. Components and items for CL

Components		Items
Corpus literacy (CL)	Understanding (U)	1. I understand what a corpus is (U1)
		2. I understand what a concordance line is (U2)
		3. A corpus contains authentic language data of a specific type (such as fiction, news, academic, and correspondence) (U3)
		4. A corpus contains written or spoken language (U4)
	Search (S)	5. I understand how to work with a corpus (S1)
		6. I know how to search corpus data using keywords (S2)
		7. I know how to search for collocates in corpus data (S3)
	Analysis (A)	8. I know how to analyse searched concordance lines (A1)
		9. I examine the words before or after the keyword in concordance lines (A2)
		10. It is useful to pay attention to the punctuation marks before and after the keyword (A3)
	Advantages (Ad)	11. I can draw conclusions about language use after searching corpus data (Ad1)
		12. I can discover language use trends after searching corpus data (Ad2)
		13. I can summarise the language use patterns after observing concordance lines (Ad3)
	Limitations (L)	14. I am aware of the limitations of using corpus data for my language learning (L1)
		15. I am aware of the limitations of using corpus data for language teaching (L2)
		16. I know what other resources to use to overcome the limitations of using corpus data (L3)

Table 4. Components and items for intention to develop corpus-based language pedagogy

Component	Items
Intention to integrate corpora into classroom teaching (ITICT)	1. A corpus is a useful data source for my language teaching (In1)
	2. I will try to use corpus data to help with my language teaching in the future (In2)
	3. I will design some corpus-based language learning tasks for my students in the future (In3)

relationship was also weak (Shieh & Wu, 2016). For a good scale, Ferketich (1991) recommended that corrected item-total correlations should range between 0.30 and 0.70. The results of our item analyses revealed very good values for all items in the six components (five for CL and one for ITICT), ranging from 0.89 to 0.55.

The next step involved conducting a reliability analysis to test the extent to which the instrument was likely to measure various components of CL consistently. We used Cronbach's alpha coefficients to evaluate the internal consistency reliability associated with scores derived from a scale. By default, any Cronbach's alpha exceeding 0.70 can be deemed acceptable.

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

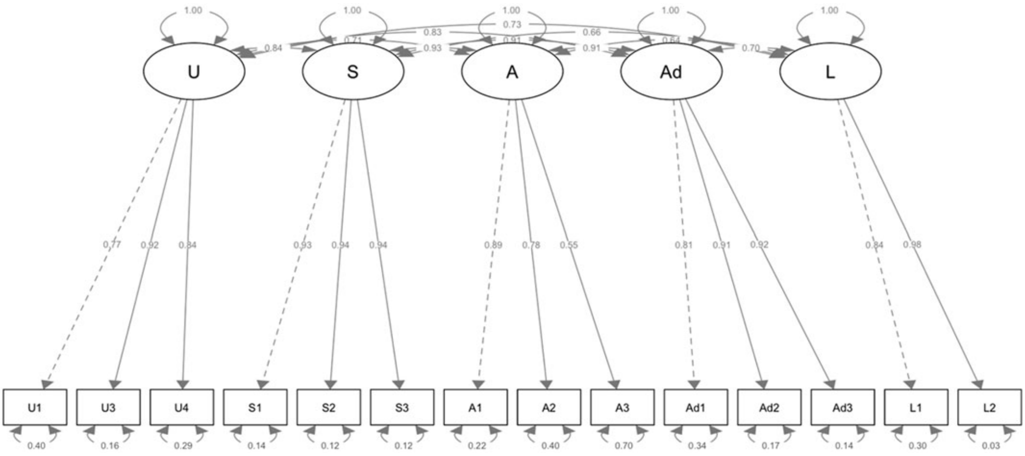


Figure 1. Factor structure of the CL scale with the five subscales

3.4.2 Confirmatory factor analysis

To assess the structural validity of the 5-factor scale for CL (see Figure 1) based on the developed items, we used the confirmatory factor analysis (CFA) function in R software (R Core Team, 2021) using Rosseel's (2012) lavaan package. For the five subscales, we included the items' respective latent subscales. To evaluate the goodness of fit of the 5-factor model, several fit indices were evaluated using chi-squared (χ^2) statistics and other goodness-of-fit indices: the comparative fit index (CFI), the Tucker–Lewis index (TLI), the root-mean-square error of approximation (RMSEA), and the standardised root-mean-square residual (SRMR). According to Hu and Bentler's (1995) recommendation, a model with CFI and TLI values > 0.90 and an RMSEA value of < 0.08 fitted well with the data. An SRMR value of < 0.08 was considered to be a good fit and value of 0.00 was a perfect fit (Hu & Bentler, 1999). The *p*-value of the χ^2 should be > 0.05. However, when the sample size is large – as in the present study – a nonsignificant χ^2 may be difficult to obtain (Barret, 2007).

3.4.3 SEM analysis

After the structural validity was tested, we proceeded to test the SEM using STATA 14 MP to predict the ITICT. Since the factor loadings were already identified by the CFA (see Comrey & Lee, 1992), we used the sum scores of the five factors as observed variables for the latent variable CL. The sum score method is especially applicable when the scales used are still exploratory (DiStefano, Zhu & Mindrila, 2009; Hair, Black, Babin, Anderson & Tatham, 2006), and exploratory studies generally find this approach acceptable (Tabachnick & Fidell, 2001). Subsequently, we used the latent variable CL to predict the latent variable regarding ITICT with its three observed variables (see Figure 2).

3.4.4 Analysis of corpus-based teaching activities

Our survey only measured the participants' perceived CL, not their actual performance. To compensate for this limitation, we will analyse some of the CL skills demonstrated in some corpus-based lesson activities designed by the participants. However, for brevity, only one corpus-based lesson will be selected and analysed. This analysis will focus on whether the designed activities reflected some of the five CL components outlined in Table 3.

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

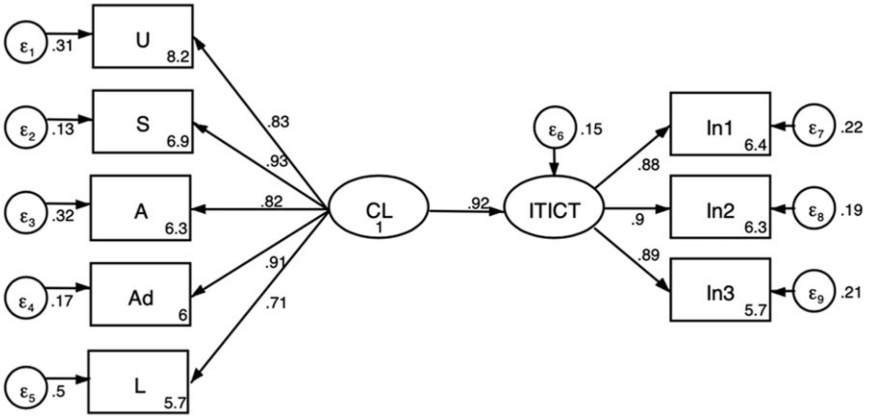


Figure 2. Structural equation model using CL as a predictor of ITICT
Note. U = Usage; S = Search; A = Analysis; Ad = Advantages; L = Limitations; ITICT = Intention to integrate corpora into classroom teaching; In = Intention.

3.4.5 Open-ended data analysis

All the answers to the two open-ended questions were collected and entered into Excel spreadsheets for the content analysis. The analysis proceeded in two steps (coding and thematic analysis) following the guidance provided by Creswell and Guetterman (2019) and Braun & Clarke, 2006). The data were coded independently by two of the authors, and the inter-code reliability reached 88%. All disputed cases were resolved through discussion and agreement, after which the codes were combined to form a number of themes.

4. Results

4.1 Likert-scale data

The results of our reliability analysis, measured by Cronbach's alpha for each subscale, indicated high reliability (0.954–0.775). As a result, we retained the 19 items for the overall scale. The reliability results for the final subscales are listed in Table 5.

To test the structural validity of the scale, we tested the 5-factor structure using all the items. This 5-factor structure was based on the five theoretical components of CL we constructed based on the work of Mukherjee (2006) and Callies (2016). This model included all the items (see Model 1 in Table 6). Model 1 had the following indices: CFI = 0.931, TLI = 0.914, and RMSEA = 0.098, with a 90% confidence interval of 0.087–0.110 and an SRMR of 0.06. Since the two items (U2 and L3) had cross-loadings of ≥ 0.45, these items had a poor fit to the model (see Brown, 2015; Tabachnick & Fidell, 2012). Accordingly, we removed these two items and ran the CFA again, which resulted in Model 2 (see Figure 1). Model 2 significantly improved the fit indices: CFI = 0.956, TLI = 0.943, and RMSEA = 0.086, with a 90% confidence interval of 0.072–0.100 and an SRMR of 0.03. This suggested that the 5-factor model without items U2 and L3 had a better fit than Model 1. To reiterate, while both models had significant χ^2 values considering the sample size (Barrett, 2007), Model 2 had better fit indices and was thus more desirable.

After validating the 5-factor structure of the CL scale, we tested whether CL would predict the ITICT using an SEM. To evaluate the model fit of the SEM, three fit indices were evaluated: CFI, TLI, and RMSEA. The SEM model (see Figure 2) had a good fit with the data, $\chi^2(19, N = 181) = 39.89$, $p < 0.05$; CFI = 0.985, TLI = 0.977; RMSEA = 0.078 (90% CI = 0.00–0.04); SRMR = 0.023. The effect of CL in predicting the ITICT in the SEM model was positive and significant ($B = 0.41$, $p < 0.001$). The full model demonstrated that 95.71% of the variance of

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Table 5. The reliability results for the subscales

Scale	Subscale	Cronbach's alpha	*N* of items
CL	1. Understanding of corpora	0.877	4
	2. Corpora search skills	0.954	3
	3. Analysis of corpus data	0.775	3
	4. Advantages of corpora	0.906	3
	5. Limitations of corpora	0.838	3
ITICT	6. Intention to integrate corpus technology into classroom teaching	0.919	3

Table 6. Comparative fit indices for alternative models of the state locus-of-hope scale

	Fit indices						
Models	χ^2	*df*	*p*	CFI	TLI	RMSEA [90% CI]	SRMR
Model 1	376.67	137	< 0.001	0.931	0.914	0.098 [0.087, 0.110]	0.06
Model 2	242.26	104	< 0.001	0.956	0.943	0.086 [0.072, 0.100]	0.03

Note. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardised root-mean-square residual.

ITICT could be predicted by all the entered observed and latent variables of CL. Specifically, the latent variable CL significantly predicted 92% of the variance of ITICT.

4.2 Corpus-based teaching activities

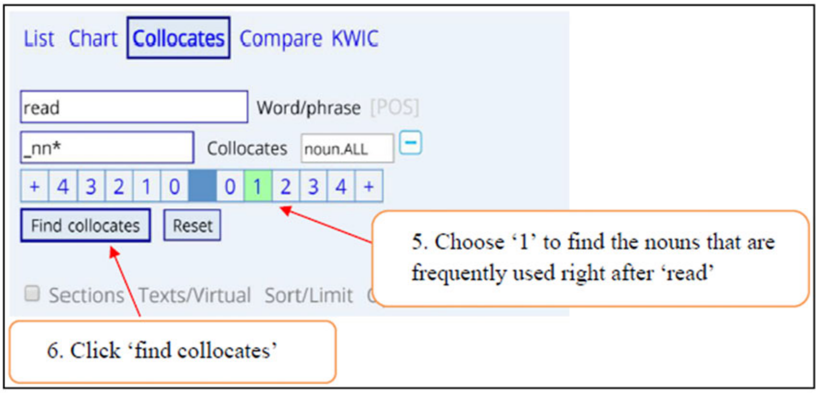
The following corpus-based teaching activities were selected from a lesson co-designed by two student teachers. The aim of the lesson was to help Chinese junior secondary 1 students differentiate three easily confused English verbs ("read", "watch", and "see"), because these can be approximately translated into the common Chinese word '看' (kan). The following analysis focuses on the five previously identified CL skills.

4.2.1 Understanding

Overall, the lesson plan (see supplementary material) featured a series of well-designed corpus-based activities using COCA and included two relevant stages: "hands-on corpus search by students" and "inductive discovery by students". The two trainees demonstrated a good understanding of corpus knowledge by integrating direct corpus consultations into their teaching steps. In addition, they designed a step-by-step COCA search guide (see Figures 3–5 for some excerpts). This demonstrated their high motivation toward adopting corpus technology as a teaching tool in the classroom.

4.2.2 Search

The two trainees designed an activity using the "collocates" function on COCA for students. After providing students with a demonstration, they are asked to follow the instructions (Figure 3) to search for nouns that collocate with "read", "watch", and "see".



read

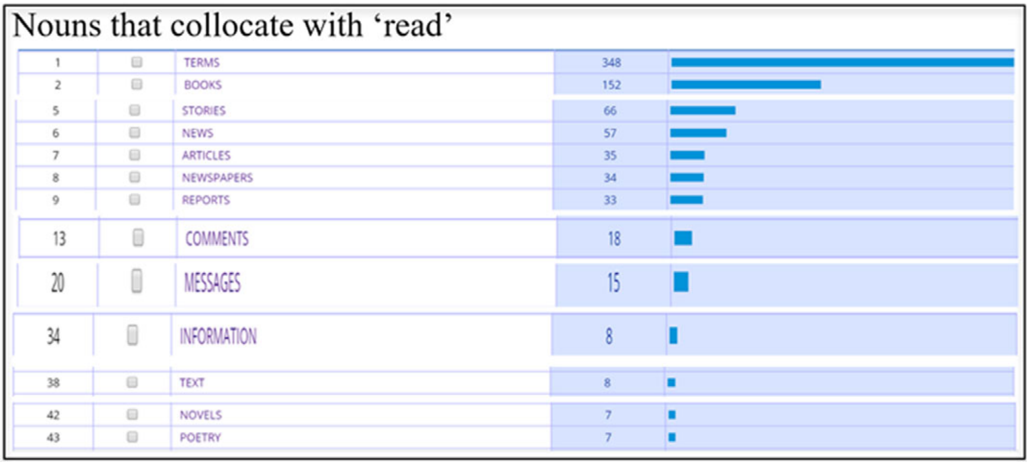
_nn*

noun.ALL

5. Choose ‘1’ to find the nouns that are frequently used right after ‘read’

6. Click ‘find collocates’

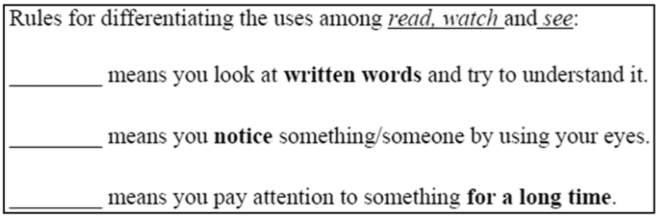
Figure 3. Excerpt from a corpus search activity



Nouns that collocate with ‘read’

1	□	TERMS	348
2	□	BOOKS	152
5	□	STORIES	66
6	□	NEWS	57
7	□	ARTICLES	35
8	□	NEWSPAPERS	34
9	□	REPORTS	33
13	□	COMMENTS	18
20	□	MESSAGES	15
34	□	INFORMATION	8
38	□	TEXT	8
42	□	NOVELS	7
43	□	POETRY	7

Figure 4. Excerpt from a corpus analysis activity



Rules for differentiating the uses among *read, watch* and *see*:

________ means you look at **written words** and try to understand it.

________ means you **notice** something/someone by using your eyes.

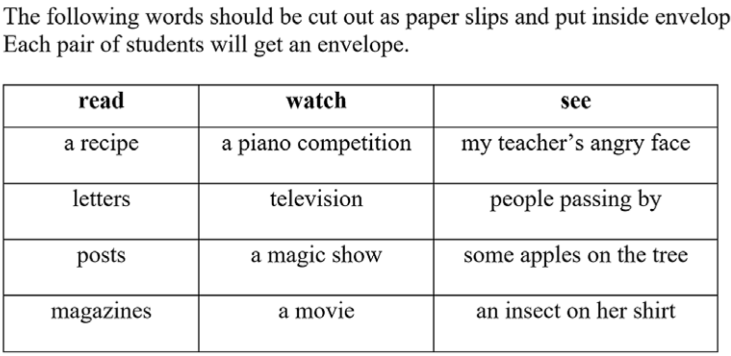
________ means you pay attention to something **for a long time**.

Figure 5. Excerpt from a summary of the use and meaning of “read”, “watch”, and “see”

4.2.3 Analysis

The two trainees understood how to encourage students to analyse the concordance results from COCA and guided them to understand the differences between the three verbs (see Figure 4). Towards the end, the students were also provided with a “fill-in-the-blanks” exercise as a way of summarising the meaning and use of the three words (see Figure 5).

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press



The following words should be cut out as paper slips and put inside envelopes. Each pair of students will get an envelope.

read	watch	see
a recipe	a piano competition	my teacher's angry face
letters	television	people passing by
posts	a magic show	some apples on the tree
magazines	a movie	an insect on her shirt

Figure 6. Excerpt from the non-corpus activity

4.2.4 Advantages and limitations

The two trainees took advantage of the collocate function of COCA when teaching collocations of the three target words in a series of coherent activities, as shown previously. In addition, they were aware of the difficulties faced by junior secondary school learners when summarising language use patterns from concordance lines. Therefore, they tried to integrate some interactive and non-corpus activities into the corpus-based lesson. For example, prior to asking students to summarise the word meaning and uses (Figure 5), they designed an activity (see Figure 6) where students needed to match nouns/noun phrases (e.g. recipe, letters, television, and teacher's angry face) with the three target verbs in the form of paper slips in pair work. The activity shown in Figure 6 demonstrates the two student teachers' awareness of the limitations of corpus data and their attempts to ameliorate the situation by scaffolding student learning with interactive, non-corpus resources, as recommended in Ma *et al.* (2021).

Similar to the studies by Ma *et al.* (2021, 2022) and Crosthwaite *et al.* (2021), the analysis of the selected lesson activities indicates that our trainees had developed a good level of CL skills. Moreover, they had a good understanding of corpus technology, tried to integrate it into their lesson plan, and could take advantage of corpus tools and search functions to develop their lesson materials.

4.3 Open-ended question data

The qualitative data collected from the two open-ended questions were analysed to identify the teacher-perceived benefits and limitations of corpora use. Using thematic analysis, some themes were generated, and the results are shown in Table 7.

4.3.1 Benefits

Five key themes emerged regarding the benefits of corpora. The first was (1) *access to authentic data*, with both teachers and learners being able to access authentic language data and see how target words/phrases are naturally used in different contexts.

The second advantage was (2) *promotion of autonomous learning.* For learners, corpora can help in discovering, summarising, or self-correcting language patterns/usage in language learning. For teachers, it can help to encourage student independent/inductive/discovery learning through hands-on corpus searches.

The third advantage was (3) *opportunity for learning and teaching collocations*, and the fourth advantage was (4) *learning/teaching difficult or easily confusing language/lexical items.* Learners can learn confusing lexical pairs and enhance their accuracy of word choice in academic writing. For teachers, corpora can provide evidence that enables them to teach confusing lexical pairs and can help them confirm their intuitive language use, as revealed by the following statement:

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Table 7. Themes drawn from open-ended question data

Aspects	Themes	Frequency
Benefits	1. Access to authentic data	37
	2. Promotion of autonomous learning	36
	3. Opportunity for learning/teaching collocations	21
	4. Learning/teaching difficult language items	15
	5. Using corpus resources for designing teaching activities	14
Limitations	1. Learning difficulty for lower-level students in summarising language patterns	25
	2. Time-consuming for learners/teachers	21
	3. Limited teacher ICT skills	9
	4. Teacher difficulty in selecting/modifying appropriate language data	7
	5. Teacher difficulty in integrating corpora with other resources	5

> I can use corpora to check my understanding before I deliver a course to students. It also enables learners to enhance the accuracy of their word choices. (Open-ended questions, data 4)

The last advantage was (5) *using corpus resources for designing teaching activities*:

> We can design appropriate vocabulary teaching activities, which is very helpful. (Open-ended questions, data 5)

The results indicated that corpora can be of great benefit to both teachers and learners, reflecting their potential to enhance language learning and teaching (to a lesser extent). The most frequently mentioned benefits were the authenticity of the corpus data and the promotion of autonomous learning through corpus use. Next came the learning/teaching of linguistic items, such as collocations, lexis, or grammar. Comparatively few teacher participants perceived the benefits of corpus resources for designing teaching activities.

4.3.2 Limitations

Five main limitations regarding the use of corpus resources for language teaching and learning emerged from the data analysis. The first was (1) *learning difficulty for lower-level students in summarising language patterns*. Language data sampled in corpora may be too difficult to enable learners to summarise language use patterns without a teacher's guidance (especially those with low proficiency), as reported as follows:

> It may be difficult for beginners to draw conclusions from their observations without guidance. Teachers should prepare well to help learners to draw conclusions, which is essential. (Open-ended questions, data 6)

The analysis of the corpus-based teaching activities designed by the two student teachers demonstrated their awareness of this limitation. Accordingly, they tried to reduce the learning difficulties of the low-level learners by designing some scaffolding activities to facilitate an inductive summary of the language patterns from using concordance lines.

The second limitation was that corpus use is (2) *time-consuming* for both learners and teachers. The third limitation was (3) *a lack of access to a computer and limited (pedagogical) ICT skills*. The fourth limitation was (4) *teachers' difficulty in selecting and modifying appropriate texts*. Teachers

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

need to filter appropriate, accurate, and grammatically accurate texts from the corpus to prepare students for lessons. The final disadvantage (5) was *teachers' difficulty in integrating corpora with other teaching resources*:

> I find it difficult to integrate corpus-based activities with other resources. (Open-ended questions, data 10)

Some of the limitations perceived by the teachers aligned with previous studies that accounted for teachers' reluctance to use corpora: it can be time-consuming for learners to analyse language data and for teachers to select appropriate language data, teachers may lack confidence and corpus-related ICT skills, and the CL may be incompatible with their already packed teaching schedules.

5. Discussion and implications

The CFA analysis validated the five key factors on which CL is built, and the SEM analysis demonstrated that teachers with higher CL tended to have stronger intentions to integrate corpora into classroom teaching. The analysis of the corpus-based teaching activities included in one selected lesson provided a snapshot of the trainees' performance data on CL skills. The analysis of teacher-perceived benefits confirmed the potential of corpora for language learning/teaching and limitations highlighted teachers' reluctance to explore corpus resources as a direct teaching tool.

5.1 Key factors for CL

Our study proposed and confirmed five subskills for CL, advancing our understanding of CL as a theoretical concept that lays the foundation for corpus teacher training. Within our research, a systematic measure of CL using both Likert-scale and open questions was also conducted, making important contributions to the measurement of CL. Built on Mukherjee's (2006) four-component CL, our results empirically validated Callies' (2016) proposal that search skills should be included in CL. The newly added *search skills* turned out to be the most important factor for acquiring CL, since the path coefficient for search skills had the highest value of 0.93 (see Figure 2). Essentially, CL is "a bundle of complex skills conceived of as the ability to use the tools and technology of corpus linguistics" (Callies, 2016: 391) to enhance language learning and teaching. The key to exploring any corpus data is to learn how to use concordances to perform various corpus searches and to generate concordance lines for observations. The *advantages* factor of using corpora also yielded a high path coefficient of 0.91. This factor was similar to the often quoted "perceived usefulness" in the TAM model (Davis, 1989), which is a key component for determining an individual's attitude towards the use of technology and their acceptance behaviour. If teachers clearly perceive the usefulness (or advantages) of corpora, this would be a driving force for them deciding to learn and make use of corpora to enhance their language teaching. The importance of the search and analysis skills is supported by the analysis of the corpus-based teaching activities designed by our trainees (see above).

The two factors of *understanding* of corpora and *analysis* of corpus data also had relatively high path coefficients (0.83 and 0.82, respectively). A basic understanding of corpora is a prerequisite, providing motivation for using corpus tools and technology. Moreover, knowing how to analyse corpus data (especially concordance lines) is an essential skill that needs to be mastered. This skill facilitates manipulating corpus tools and extrapolating meaning and usage from the authentic language examples provided by the corpora. The analysis of the lesson designed by our trainees also demonstrated the importance of understanding and analysing corpus data with regard to developing an essential CL.

The coefficient for the limitations of using corpora was slightly lower than for the first four factors of CL at 0.71; however, it remained statistically significant ($p < 0.001$). No technology

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

is perfect (including corpora), and understanding the limitations of corpora was one of the most challenging perspectives in relation to CL development (Zareva, 2017). In addition, greater awareness of the limitations was associated with stronger CL development among teacher trainees (Heather & Helt, 2012). In this sense, we were pleased to establish that some of our trainees were aware of the limitations of corpus data and tried to use other means to compensate for these limitations when designing their corpus-based lesson materials.

5.2 The relationship between CL and teachers' intention to integrate corpora into classroom teaching

Previous research investigating teacher perceptions of corpora and teacher attitudes towards incorporating corpora into their classroom teaching has typically yielded contradictory pictures. While the majority of teachers or student teachers clearly acknowledge the great potential that corpora have for language learning, they are much less willing to use corpora in their classroom teaching (Breyer, 2009; Naismith, 2017; Zareva, 2017). The current study took a step further by investigating teachers' CL, which clearly demonstrated that CL leads to teachers' intention to use corpora in classroom teaching. In other words, those teachers or student teachers with a higher level of CL are more likely to form a positive intention of incorporating corpora into classroom use. Given the high importance of search skills, more relevant training activities should be provided to help teachers gain familiarity with the various corpus search functions available on popular corpus websites. Moreover, enhancing their corpus search skills may also increase their confidence in working with corpora (Heather & Helt, 2012; Naismith, 2017).

5.3 Teacher-perceived benefits and limitations of exploiting corpora in classroom teaching

In this study, the benefits that participants perceived in corpus use in language classrooms were also identified. Participant responses concentrated on five main themes: access to authentic data, learning/teaching collocations, promotion of autonomous learning, learning/teaching difficult language items, and using corpus resources for designing teaching activities. These teacher perceptions aligned with the reported benefits of corpora for language education (Heather & Helt, 2012; Leńko-Szymańska, 2017; McEnery & Xiao, 2011). For example, corpora are perceived to offer a considerable advantage in addressing learners' collocation difficulties/errors (Fang, Ma & Yan, 2021; McEnery & Xiao, 2011; Tsai, 2019). From a pedagogical perspective, another important advantage of using corpora in language teaching is by helping teachers to promote a more learner-centred, autonomous approach to language learning and teaching (Boulton, 2017). An analysis of these teacher-perceived benefits confirmed the great potential that corpora have for language learning. Finally, the results indicated that teachers were inclined to view corpora more as a learning tool rather than as a classroom teaching tool. The reasons for this may be partly due to the perceived limitations of corpora usage.

The results revealed five teacher-perceived limitations. One new finding that has rarely been reported was the *teachers' difficulty in integrating corpora with other resources*. Again, this could explain why teachers preferred to use corpora as a learning rather than teaching tool (Breyer, 2009; Leńko-Szymańska, 2017; Naismith, 2017). The literature also presented various reasons why teachers are reluctant to apply corpora in their teaching. These included frustration with the technical problems associated with corpora (Breyer, 2009; Farr, 2008; Naismith, 2017), the heavy workload involved (Leńko-Szymańska, 2017; Lin & Lee, 2015), and a lack of corpora training (Boulton, 2017; Tribble, 2015). The current study extended this understanding by revealing a new reason why teachers are reluctant to employ corpora into classroom teaching – the difficulty of integrating corpora with other resources.

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

5.4 Pedagogical implications

Based on our results, we propose several pedagogical implications for corpus linguists, corpus website developers, and teacher educators to help them provide successful CL training for teachers or students.

The first concerns what corpus skills should be the focus in CL training. Since the understanding of corpora is usually the first topic to be covered in corpus teacher training, it is assumed that this understanding can be relatively easily acquired. Further, it has been reported that teachers and student teachers are able to identify the benefits of corpora use after their training (e.g. Farr, 2008; Heather & Helt, 2012; Naismith, 2017; Zareva, 2017), which are mainly due to the advantages of corpus linguistics itself. Therefore, this aspect of training would also not be too challenging. The remaining three subfactors of analysis, awareness of limitations, and especially search skills (which are core components of the training) are worthy of greater attention.

The second concern is the sequence of introducing corpus websites to participants. The initial learning of corpora may be quite challenging for teachers and students, as this includes the functions and skills needed to perform essential concordance searches of corpus data, including searches for keywords or collocates (Leńko-Szymańska, 2017; Naismith, 2017; Zareva, 2017). Although the keyword search function is similar for many corpora concordancers, some online corpus websites have their own unique operating systems and different search functions (e.g. Lextutor, BNC, and COCA). It is suggested that only one corpus website is focused on at a time. Only after the participants have become familiar with the key search functions of a particular corpus should other (similar or different) corpus websites be introduced to participants sequentially.

Third, to facilitate teachers and learners using corpora, corpus website developers may consider using similar search interfaces and simplifying search syntax. This was recently conducted for the COCA website, where many complicated search syntaxes (previously only understood by corpus linguists) were simplified and reduced to a limited set of straightforward search formulae. This practice is desirable and presents a good example for other corpus websites to follow to encourage increasing numbers of teacher and learner users.

Fourth, corpus analysis skills should also be given sufficient attention in the training. Closely related to search skills, corpus analysis skills may be considered demanding by teachers and student teachers. Zareva (2017) reported that teacher trainees considered the analysis of corpus data time-consuming and that it was not easy to summarise patterns of language use. Further, they encountered difficulties in understanding the information in tables and charts. For these reasons, we suggest that more hands-on corpus practice involving corpus analysis should be provided during training; for example, both Breyer (2009) and Farr (2008) emphasised that more opportunities should be provided for trainees to interact with corpora as learners. From the learner perspective, Heather and Helt (2012) proposed that teacher trainees should learn how to "organize and present concordance data in ways that lead more clearly to autonomous learning for their students" (p. 436).

Finally, helping teachers to become aware of the limitations of corpora usage and to identify alternative solutions are important aspects of CL training. The implication was that the more aware of limitations teachers are, the more likely they are to develop methods of compensating for any limitations. Hence, they would make better use of corpora to facilitate student learning. This implies that more specific training on how to compensate for corpus limitations is needed to develop rounded CL for teachers. In addition, previous research suggests that understanding corpus limitations may be indicative of a high level of CL (Heather & Helt, 2012). Understanding teacher-perceived limitations pertaining to corpora use may serve three purposes. First, knowing about these limitations may help corpus linguists and concordancer developers to understand the difficulties teachers experience with corpora usage. This would allow them to develop ideas for improving corpus tools and to align them with teacher needs. Second, educators

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

providing corpus-based training for teachers should consider these limitations and design effective instructional activities/procedures to overcome them. Third, raising teacher awareness of the limitations means enhancing their metacognition and helping them to manage and plan effective corpus-based teaching.

6. Conclusions and limitations

Adopting a SEM approach, this research established five key components of CL that account for teachers' intention to integrate corpora into their classroom teaching. Although it is commonly agreed that CL should be the focus of CL teacher training, researchers have tended to interpret and select rather idiosyncratically the subskills that should be included in CL training. Our research clearly established the five factors that comprise CL (understanding, search, analysis, and the advantages and limitations of corpora); therefore, successful training should hinge on adequate training that involves these five factors. Moreover, since it is relatively easy to learn about (and understand) the advantages of corpora, greater effort should be made to provide training in the three subfactors: analysis, limitations, and especially search skills of corpora (which is the foundation of corpus-based training). Finally, understanding the limitations of corpora can raise teachers' metacognition, especially for devising alternative solutions to make better use of corpora in their classroom teaching.

This study has two limitations: self-report data and lack of concrete tests to assess participant use of corpora in their teaching. As self-report data can be subjective, the Likert-type scale measures of CL might differ from those of CL performance data. Hence, future studies can develop and psychometrically validate CL test items to measure teachers' CL performance data more objectively and accurately. Additional psychometrically validated objective measures include teacher use of corpus search functions, analysis of concordance lines, and how they integrate CL into designing and using suitable teaching materials to facilitate their classroom teaching.

Supplementary material. To view supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344022000180

Acknowledgements. The article was supported by two CRAC projects (Ref: 03AAB; 04A32) at The Education University of Hong Kong.

Ethical statement and competing interests. Accessing the data archive is determined by the funding body of the research, namely The Education University of Hong Kong. All participants participated voluntarily in this research and appropriate ethical procedure was followed to obtain participant consent. The authors declare no competing interests.

References

Ackerley, K. (2021) Exploiting a genre-specific corpus in ESP writing: Students' preferences and strategies. In Charles, M. & Frankenberg-Garcia, A. (eds.), *Corpora in ESP/EAP writing instruction: Preparation, exploitation, analysis*. New York: Routledge, 78–99. https://doi.org/10.4324/9781003001966-4-7

Barrett, P. (2007) Structural equation modelling: Adjudging model fit. *Personality and Individual differences*, 42(5): 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Boulton, A. (2017) Corpora in language teaching and learning. *Language Teaching*, 50(4): 483–506. https://doi.org/10.1017/S0261444817000167

Boulton, A. & Cobb, T. (2017) Corpus use in language learning: A meta-analysis. *Language Learning*, 67(2): 348–393. https://doi.org/10.1111/lang.12224

Boulton, A. & Vyatkina, N. (2021) Thirty years of data-driven learning: Taking stock and charting new directions over time. Language Learning *& Technology*, 25(3): 66–89. https://doi.org/10125/73450

Braun, V. & Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2): 77–101. https://doi.org/10.1191/1478088706qp063oa

Breyer, Y. (2009) Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, 22(2): 153–172. https://doi.org/10.1080/09588220902778328

Brown, T. A. (2015) *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

Çalışkan, G. & Kuru Gönen, S. İ. (2018) Training teachers on corpus-based language pedagogy: Perceptions on using concordance lines in vocabulary instruction. *Journal of Language and Linguistic Studies*, 14(4): 190–210.

Callies, M. (2016) Towards corpus literacy in foreign language teacher education: Using corpora to examine the variability of reporting verbs in English. In Kreyer, R., Schaub, S. & Güldenring, B. A. (eds.), *Angewandte Linguistik in Schule und Hochschule [Applied linguistics in secondary school and at university]*. Frankfurt am Main: Peter Lang, 391–415.

Callies, M. (2019) Integrating corpus literacy into language teacher education: The case of learner corpora. In Götz, S. & Mukherjee, J. (eds.), *Learner corpora and language teaching*. Amsterdam: John Benjamins, 245–263. https://doi.org/10.1075/scl.92.12cal

Chang, J.-Y. (2014) The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, 26(2): 243–259. https://doi.org/10.1017/S0958344014000056

Charles, M. (2012) 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31(2): 93–102. https://doi.org/10.1016/j.esp.2011.12.003

Charles, M. (2014) Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, 35: 30–40. https://doi.org/10.1016/j.esp.2013.11.004

Charles, M. (2018) Using do-it-yourself corpora in EAP: A tailor-made resource for teachers and students. *Journal of Teaching English for Specific and Academic Purposes*, 6(2): 217–224. https://doi.org/10.22190/jtesap1802217c

Charles, M. & Hadley, G. (2022) Autonomous corpus use by graduate students: A long-term trend study (2009–2017). *Journal of English for Academic Purposes*, 56: Article 101095. https://doi.org/10.1016/j.jeap.2022.101095

Chen, H.-J. H. (2011) Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, 24(1): 59–76. https://doi.org/10.1080/09588221.2010.526945

Chen, M., Flowerdew, J. & Anthony, L. (2019) Introducing in-service English language teachers to data-driven learning for academic writing. *System*, 87, 102148. https://doi.org/10.1016/j.system.2019.102148

Comrey, A. L. & Lee, H. B. (1992) *A first course in factor analysis* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Creswell, J. W. & Guetterman, T. C. (2019) *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). New York: Pearson Education.

Crosthwaite, P., Luciana, & Wijaya, D. (2021). Exploring language teachers' lesson planning for corpus-based language teaching: a focus on developing TPACK for corpora and DDL. *Computer Assisted Language Learning*, 1–29. https://doi.org/10.1080/09588221.2021.1995001

Davis, F. D. (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3): 319–340. https://doi.org/10.2307/249008

DiStefano, C., Zhu, M. & Mindrila, D. (2009) Understanding and using factor scores: Considerations for the applied researcher. Practical Assessment, *Research, and Evaluation*, 14(1): 20. https://doi.org/10.7275/da8t-4g52

Ebrahimi, A. & Faghih, E. (2017) Integrating corpus linguistics into online language teacher education programs. *ReCALL*, 29(1): 120–135. https://doi.org/10.1017/S0958344016000070

Fang, L., Ma, Q. & Yan, J. (2021) The effectiveness of corpus-based training on collocation use in L2 writing for Chinese senior secondary school students. *Journal of China Computer-Assisted Language Learning*, 1(1): 80–109. https://doi.org/10.1515/jccall-2021-2004

Farr, F. (2008) Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness*, 17(1): 25–43. https://doi.org/10.2167/la414.0

Ferketich, S. (1991) Focus on psychometrics: Aspects of item analysis. *Research in Nursing & Health*, 14(2): 165–168. https://doi.org/10.1002/nur.4770140211

Fitzgerald, A. T. D. (2018) *A new paradigm for open data-driven language learning systems design in higher education*. Concordia University, PhD thesis.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. & Tatham, R. L. (2006) *Multivariate data analysis*. Upper Saddle River: Prentice Hall.

Heather, J. & Helt, M. (2012) Evaluating corpus literacy training for pre-service language teachers: Six case studies. *Journal of Technology and Teacher Education*, 20(4): 415–440.

Hu, L. & Bentler, P. M. (1995) Evaluating model fit. In Hoyle, R. H. (ed.), *Structural equation modeling: Concepts, issues, and applications*. Sage Publications, 76–99.

Hu, L. & Bentler, P. M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1): 1—55. https://doi.org/10.1080/10705519909540118

Johns, T. (1991) Should you be persuaded: Two examples of data-driven learning. In Johns, T. & King, P. (eds.), Classroom concordancing. *English Language Research Journal*, 4: 1–16.

Latif, M. M. M. A. (2021) Corpus literacy instruction in language teacher education: Investigating Arab EFL student teachers' immediate beliefs and long-term practices. *ReCALL*, 33(1): 34–48. https://doi.org/10.1017/S0958344020000129

Lee, H., Warschauer, M. & Lee, J. H. (2017) The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, 21(2): 32–51.

Leech, G. (1997) Teaching and language corpora: A convergence. In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (eds.), *Teaching and language corpora*. London: Longman, 1–23. https://doi.org/10.4324/9781315842677-1

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Leńko-Szymańska, A. (2014) Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL*, 26(2): 260–278. https://doi.org/10.1017/S095834401400010X

Leńko-Szymańska, A. (2017) Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology*, 21(3): 217–241.

Leńko-Szymańska, A. & Boulton, A. (eds.). (2015) *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.69

Lin, M. H. & Lee, J.-Y. (2015) Data-driven learning: Changing the teaching of grammar in EFL classes. *ELT Journal*, 69(3): 264–274. https://doi.org/10.1093/elt/ccv010

Ma, Q., Tang, J. & Lin, S. (2021) The development of corpus-based language pedagogy for TESOL teachers: A two-step training approach facilitated by online collaboration. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2021.1895225

Ma, Q., Yuan, R., Cheung, L. M. E. & Yang, J. (2022) Teacher paths for developing corpus-based language pedagogy: A case study. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2022.2040537

McEnery, T. & Xiao, R. (2011) What corpora can offer in language teaching and learning. In Hinkel, E. (ed.), *Handbook of research in second language teaching and learning* (Vol. 2). New York: Routledge, 364–380.

Mukherjee, J. (2006) Corpus linguistics and language pedagogy: The state of the art – and beyond. In Braun, S., Kohn, K. & Mukherjee, J. (eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods*. Frankfurt am Main: Peter Lang, 5–24.

Naismith, B. (2017) Integrating corpus tools on intensive CELTA courses. *ELT Journal*, 71(3): 273–283. https://doi.org/10.1093/elt/ccw076

O'Keeffe, A. & Farr, F. (2003) Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly*, 37(3): 389–418. https://doi.org/10.2307/3588397

Özbay, A. & Kayaoğlu, M. N. (2015) EFL teacher's reflections towards the use of computerized corpora as a teaching tool in their classrooms. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi*, 44(1): 85–104. https://doi.org/10.14812/cuefd.54367

Pérez-Paredes, P. (2019) The pedagogic advantage of teenage corpora for secondary school learners. In Crosthwaite, P. (ed.), *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Abingdon: Routledge, 67–87. https://doi.org/10.4324/9780429425899-5

Pérez-Paredes, P. (2022) A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, 35(1–2): 36–61. https://doi.org/10.1080/09588221.2019.1667832

Pérez-Paredes, P. & Mark, G. (eds.) (2021) *Beyond concordance lines: Corpora in language education*. Amsterdam: John Benjamins.

Pérez-Paredes, P., Ordoñana Guillamón, C. & Aguado Jiménez, P. (2018) Language teachers' perceptions on the use of OER language processing technologies in MALL. *Computer Assisted Language Learning*, 31(5–6): 522–545. https://doi.org/10.1080/09588221.2017.1418754

Poole, R. (2022) "Corpus can be tricky": Revisiting teacher attitudes towards corpus-aided language learning and teaching. *Computer Assisted Language Learning*, 35(7): 1620–1641. https://doi.org/10.1080/09588221.2020.1825095

R Core Team. (2021) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/

Rodgers, O., Chambers, A. & Le Baron-Earle, F. (2011) Corpora in the LSP classroom: A learner-centred corpus of French for biotechnologists. *International Journal of Corpus Linguistics*, 16(3): 391–411. https://doi.org/10.1075/ijcl.16.3.06rod

Rosseel, Y. (2012) lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2): 1–36. https://doi.org/10.18637/jss.v048.i02

Schmidt, N. (2022) Unpacking second language writing teacher knowledge through corpus-based pedagogy training. *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344022000106

Shieh, J.-I. & Wu, H.-H. (2016) Measures of consistency for DEMATEL method. *Communications in Statistics – Simulation and Computation*, 45(3): 781–790. https://doi.org/10.1080/03610918.2013.875564

Shulman, L. (1987) Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1): 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tabachnick, B. G. & Fidell, L. S. (2001) Principal components and factor analysis. In Tabachnick, B. G. & Fidell, L. S. (eds.), *Using multivariate statistics* (4th ed.). Boston: Allyn & Bacon, 582–633.

Tabachnick, B. G. & Fidell, L. S. (2012) *Using multivariate statistics* (6th ed.). San Francisco: Pearson.

Tribble, C. (2015) Teaching and language corpora: Perspectives from a personal journey. In Leńko-Szymańska, A. & Boulton, A. (eds.), *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, 37–62. https://doi.org/10.1075/scl.69.03tri

Tsai, K.-J. (2019) Corpora and dictionaries as learning aids: Inductive versus deductive approaches to constructing vocabulary knowledge. *Computer Assisted Language Learning*, 32(8): 805–826. https://doi.org/10.1080/09588221.2018.1527366

Vyatkina, N. & Boulton, A. (2017) Corpora in language learning and teaching. *Language Learning & Technology*, 21(3): 1–8. https://doi.org/10125/44621

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press

Yeh, Y., Liou, H. C. & Li, Y. H. (2007) Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning*, 20(2): 131–152. https://doi.org/10.1080/09588220701331451

Yoon, H. & Hirvela, A. (2004) ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4): 257–283. https://doi.org/10.1016/j.jslw.2004.06.002

Zareva, A. (2017) Incorporating corpus literacy skills into TESOL teacher training. *ELT Journal*, 71(1): 69–79. https://doi.org/10.1093/elt/ccw045

About the authors

Qing Ma is an associate professor at the Department of Linguistics and Modern Language Studies, The Education University of Hong Kong. Her main research interests include second language vocabulary acquisition, corpus linguistics, computer-assisted language learning and mobile-assisted language learning. Her current research focuses on how to theorise and validate empirically a corpus-based language pedagogy.

Ming Ming Chiu is Chair Professor of Analytics and Diversity (Honor) in the Department of Special Education and Counselling and Director of the Assessment Research Centre at The Education University of Hong Kong. He studies automatic statistical analyses, inequalities, culture, and learning in 65 countries. His research interests include learning analytics, group processes, inequality, corruption, and online sexual predators.

Shanru Lin is a research assistant at the Department of Linguistics and Modern Language Studies, The Education University of Hong Kong. Her main research interests include data-driven learning, technology-enhanced learning, corpus linguistics, mobile-assisted language learning, personalised learning, vocabulary learning.

Norman B. Mendoza is a postdoctoral fellow at The Education University of Hong Kong, in the Department of Curriculum and Instruction. His research interests are in assessment, motivation, and psychology in the school and educational context.

Author ORCiD. iD Qing Ma, https://orcid.org/0000-0003-3125-3513
Author ORCiD. iD Ming Ming Chiu, https://orcid.org/0000-0002-5721-1971
Author ORCiD. iD Shanru Lin, https://orcid.org/0000-0002-1439-2514
Author ORCiD. iD Norman B. Mendoza, https://orcid.org/0000-0003-0344-0709

https://doi.org/10.1017/S0958344022000180 Published online by Cambridge University Press