A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015

Pascual Pérez-Paredes
University of Cambridge

Author contact details (Main corresponding author)

Pascual Pérez-Paredes
Faculty of Education
University of Cambridge
184 Hills Road
CB2 8PQ
Cambridge, UK

e-mail: pfp23@cam.ac.uk

Word count of the manuscript: 9,113 (incl. references), 10,137 (incl. Appendices).

Abstract

This research uses the theoretical framework of CALL normalisation developed by Bax (2003) and Chambers & Bax (2006) to offer a systematic review (Gough et al., 2012) of the uses and spread of data-driven learning (DDL) and corpora in language learning and teaching across five major CALL-related journals during the 2011–2015 period. DDL research represented 4.2% of all published papers on CALL during this time frame. The main focus of research was found to be the use of concordancing and collocations when developing university students' writing skills. Contrary to previous research, access to technology was not identified as an impeding factor for normalisation. Syllabus integration and a lack of contribution from language teachers other than researchers emerged as threats to the normalisation of corpora use. Further theorisation is needed if DDL and corpora are to expand their influence on mainstream second language education.

Keywords: DDL, corpora, language education, language learning, language teaching, normalisation, CALL, systematic review

1. Introduction

According to Davies, Otto & Rüschoff (2013: 34), we have already entered a phase of CALL where digital tools for learning have become integrated elements "both in the real world and also in foreign language syllabuses". A recent survey of predominantly higher education (44%) and secondary (23%) language teachers (n=230) in Spain and the UK (Pérez-Paredes et al., 2018) corroborates this claim, given that around 70% of the teachers surveyed use either online platforms or web-based services in their everyday teaching. This survey, however, found that only a small number of these teachers were familiar with L1 corpora or learner corpora when teaching languages. Based on corpus linguistics research methods, data-driven learning (DDL) has been used in language classrooms worldwide with varying degrees of success. The literature ranges from endorsing the benefits of DDL and proclaiming its superiority over other learning approaches (Mizumoto & Chujo, 2015) to voicing learners' foremost problems when confronted with DDL (Luo, 2016). The metanalyses performed by Boulton & Cobb (2017) and Lee et al. (2018) demonstrate that DDL studies yield medium to high effect sizes in both within-group and between-group designs. However, the spread of DDL and language corpora in language learning is limited.

Normalisation refers to "the stage when the technology becomes invisible, embedded in everyday practice and hence normalised" (Bax, 2003: 23). We will draw on Bax (2003) and Chambers & Bax's (2006) notion of normalisation as the theoretical framework enabling us to understand how DDL and corpora uses have been introduced into different language learning contexts worldwide and how technological and pedagogical DDL perspectives interact in the practices under analysis. In this paper, we will examine research that has sought to make DDL and/or corpora more readily available to language learners during the

five-year period from 2011 to 2015, and we will attempt to identify the factors that impede or promote their adoption in language education.

2. DDL, CALL and normalisation

2.1 DDL and CALL

DDL entails language learners working with written or spoken data, resulting in "increased language sensitivity, noticing, induction, and ability to work with authentic data" (Boulton & Cobb, 2017: 349). Although DDL implies the use of computers and computer software such as concordancers, DDL is not always included in descriptions of CALL. Gimeno-Sanz (2016) does not list DDL or corpora in her inventory of technologies and skills; and in Grgurović, Chapelle & Shelley's (2013) meta-analysis of effectiveness studies on computer technology-supported language learning from1970 to 2006, no reference is made to either DDL, language corpora or the use of the web as a corpus. Steel & Levy (2013) surveyed 587 undergraduate language learners at an Australian University about their technology use, but no references to corpora were made at any point in this paper. In Thomas, Reinders & Warschauer (2014), corpora and DDL are only briefly mentioned in the chapter written by Davies, Otto & Rüschoff (2014). In contrast, Golonka et al. (2014: 72), include "corpus" as a type of individual study tool. Corpora provide access to rich, authentic input; enable broad access to linguistic data; and promote data-driven inductive learning. The aforementioned authors also argue that corpus linguists tend to overstate the claims that corpora can affect language learning, maintaining (p. 78) that these claims are "stronger than actual evidence for their efficacy".

It seems that DDL is not currently perceived as a major area of practice in CALL-related research. The reasons are varied, complex and well beyond the scope of this paper. DDL has its roots in research led by pioneering linguists who, at the time, regarded the use of corpora as an extension of their language-oriented research (Pérez-Paredes, 2010), a point later taken up by Vyatkina & Boulton (2017). This fact may have prevented DDL from becoming mainstream in a foreign language education field dominated by second language acquisition (SLA) and foreign language teaching (FLT) applied linguists in the 80s and 90s, when focus on form was not part of the L2 research agenda. In fact, corpora and DDL did not feature in the *Key Concepts in ELT* section of the *ELT Journal* until 2011 (Huang, 2011). It may seem circumstantial, but this glossary started in 1993 with the intention to assist *ELT Journal* readers in developing an appreciation of central ideas in ELT, that is, years after Johns' (1990) first publications on DDL. It took almost two decades for DDL to become one of these central ideas, together with scaffolding (1994), universal grammar (1995), computer-mediated communication (2002) and, to name another entry, blended learning (2010).

Those seeking to spread the benefits of DDL remained in the corpus linguistics *camp* (Sinclair, 2003). As a consequence, we often find that, outside the corpus linguistics literature, corpora have yet to achieve mainstream status (Braun, 2005; Pérez-Paredes, 2010; Boulton & Pérez-Paredes, 2014). Römer (2006: 129) stated over a decade ago that "a lot still remains to be done before […] we can say that corpora have actually arrived in language pedagogy", whereas Tribble (2008) identified the lack of a clear pathway for teachers into classroom corpus use. Similarly, Pérez-Paredes (2010) voiced the need for corpora that take into account the learning context in which they are used, and suggested the development of a so-called *feasibility* scenario where language teachers and material developers can go beyond

the mere adaptation of research-oriented corpus resources in the language classroom and in higher education (HE) settings.

However, DDL is trying to meet the needs of an ever-increasing number of learning contexts. Boulton & Pérez-Paredes (2014) highlighted the fact that the DDL focus is switching from corpus linguistics to language pedagogy, and that the emphasis is increasingly on L2 users and less on the technology itself. Boulton & Cobb's (2017) meta-analysis sought to elucidate whether positive learning outcomes stem from DDL by synthesising quantitative results in the form of effect sizes. Their search included the keywords (p. 355) *corpus*, *corpora*, *data-driven*, *DDL*, *Johns*, *concordancer*, *concordance and concordancing* in the context of language learning. The analysis included 64 empirical DDL studies published between 1989 and 2014. The authors concluded that (a) the effect sizes in both the within-group and between-group designs were high and increased during the 2011–2014 period when compared with the 1990–2005 and 2016–2010 periods; (b) higher effects were found in ranked journals such as *Computer Assisted Language Learning*, *Language Learning*, *Language Learning & Technology*, *ReCALL*, and *System* (a total of 25 papers); (c) large effect sizes were observed under laboratory-like conditions and in regular classrooms; and (d) larger effect sizes were found when learners used a concordancer or other types of software compared to paper-based DDL. Lee et al. (2018) observed an overall medium effect on L2 vocabulary learning in both the short and long term in direct DDL studies. In-depth vocabulary knowledge was associated with a larger effect size.

2.2 Normalisation

The notion of normalisation (Bax, 2003) has attracted the attention of researchers as it can enhance understanding of CALL's spread and uptake (Sun & Ye, 2006). In Bax (2003), we

find an analysis of the history behind CALL which puts forward three "approaches" that can replace the "phases" in Warshauer (2000). These approaches – Restricted, Open and Integrated CALL – account for different sets of theories addressing learning, software, activity types and teachers' roles. In Integrated CALL, technology is invisible, "taken for granted in everyday life" and ceases "to exist as a separate concept and field for discussion" (p. 23). Bax maintains that this phase can be reached once computers are used as an "integral part" (p. 24) of lessons and are not at the centre of them. According to this author, Open CALL approaches, already in place over a decade ago, favour a more communicatively oriented use of computers, although institutional and attitudinal problems have and continue to see students in instructed second or foreign language programmes blocked from benefiting from the range of resources available. Later, Bax (2011) (re)-defined normalisation as the stage when technology reaches "its fullest possible effectiveness in language education" and becomes "a valuable element in the language learning process" (p. 1).

Chambers & Bax (2006) studied the normalisation of CALL in two HE settings and looked at logistics, stakeholders' conceptions, knowledge and abilities, integration of syllabus and software, and training, development and support. The authors highlight that it is important to focus on more than one single factor influencing the use of CALL in order to "take account of the ecological complexity of the whole context in each case" (p.477). They believe that this complexity rules out the viability of technological one-shot solutions and conclude that, out of the 11 factors identified, syllabus integration is the one factor that should be addressed before normalisation of CALL can actually take place. An alternative vison was put forward by Gimeno-Sanz (2016), who prefers to define contemporary CALL practices as atomised CALL, whereby CALL practitioners have moved away from structured all-in-one content. As

for DDL, experts agree that their field is in the making (Boulton & Pérez-Paredes, 2014; Vyatkina & Boulton, 2017).

In this paper, we set out to research how the normalisation factors studied by Chambers & Bax (2006) explain the uses and spread of DDL and corpora in language learning and teaching. Our research questions are: (1) How much DDL and corpus research is published in the wider field of CALL?; (2) What is the focus of DDL research?; and (3) What role do the normalisation factors studied by Chambers & Bax (2006) play in the published research? We will focus on papers from 5 research journals in the 2011–2015 period. We will look at how logistics, stakeholders' conceptions, knowledge and abilities, integration of syllabus and software, and training, development and support are discussed in our body of research.

3. Materials and methods

Our research follows a systematic review strategy (Gough et al., 2012) that sets out to answer the questions outlined in Section 2. The scope of our review is framed by the notion of normalisation (Bax, 2003; Chambers & Bax, 2006) and we will examine how the factors impeding and facilitating normalisation (Chambers & Bax, 2006) manifest themselves across the papers under analysis. Following the guidelines in Gough et al. (2012: 74), we will present a synthesis of the literature that explores relevant findings from the set of analysed research papers so as to understand how DDL and corpora are used and normalised in CALL research based on the "tentative assumptions and concepts that emerge from the data".

3.1 Search strategy: journals and screened papers

Research papers dealing with either direct or indirect uses[1] (Römer, 2006) of DDL, L1 corpora or learner language corpora for language learning and teaching published in five top research journals in the field of CALL[2] during the 2011–2015 period were examined. The decision to look at this period was prompted by research stressing the need for further empirical research in the area (Boulton, 2008; Pérez-Paredes, 2010). The review was carried out between 2016 and 2018.The journals used in our analysis were *Computer Assisted Language Learning* (*CALL*) (1.14); *CALICO Journal*; *Language, Learning & Technology* (*LLT*) (1.12); *ReCALL* (1.12); and *System* (0.98). The figure in brackets is the mean of the SCImago Journal Ranking (SJR) impact factor during the 5-year period for each journal. No

[1] Direct uses entail learners and teachers consulting corpora while indirect uses encompass corpus findings used by material writers, lexicographers and teachers preparing their own activities (DDL or otherwise) using insights from corpus linguistics.
[2] Although all five journals publish CALL and CALL-related research, their scope is not necessarily limited to or constrained by CALL research exclusively.

SJR factor was available for *CALICO Journal*. Three are UK-based journals (*CALL*, *ReCALL* and *System*) and two are from the US (*CALICO Journal* and *LLT*).

Only full original research papers were included in the analysis. Book reviews and other journal sections were not considered. In total, 759 full original research papers were published in the five aforementioned journals between 2011 and 2015. An initial search of the keywords *corpus*, *corpora*, *DDL*, *data-driven learning* and *corpus-based* returned 37 potentially relevant papers (Appendix 1). After close examination, five of these papers [IDs 16, 19, 20, 24 and 33] were excluded from the final pool as the use of DDL or corpora in the language classroom we not among their aims or they were used as a research method to tap into research questions unrelated to the aim of our review.

3.2 Analysis of the research papers

For each analysed paper, we annotated the focus of the research, the research questions, and the different aspects that either explicitly or implicitly make reference to what Chambers & Bax (2006) describe as impeding factors for normalisation, that is, (a) logistics; (b) stakeholders' conceptions, knowledge and abilities; (c) syllabus and software integration; and (d) training, development and support. This information was captured in a matrix which was later used to structure and facilitate the "analysis of themes and trends across all the studies being addressed" (Gough et al., 2012: 136). Figure 1 shows a simplified version of the annotation for paper ID 13 (Wood, 2011).

Coding categories	**Logistics**	○ Quick Assist software ○ Software specifications: Use of Wikipedia and online dictionaries ○ The HE institution failed to provide support when setting up a server
	Stakeholders´ conception	○ Research method: use of interviews ○ General positive reaction to the use of the software for reading comprehension
	Stakeholders´ knowledge	○ Reading skill ○ Morphology ○ Definitions ○ Collocations
	Stakeholders´ abilities	○ Segmentation ○ Identification of constituents
	Syllabus integration	○ Different expectations: developers vs language instructors ○ No explicit discussion of integrating DDL in the curriculum/syllabus
	Training	○ Usefulness of the software largely dependant on the learners´ language competence level
	Learners	○ Higher education students ○ Learners of German

Figure 1. A simplified version of the annotation for one of the analysed papers (ID 13).

Sketch Engine multiword keyword analysis (uKilgariff, 2012) was used to uncover the scope and themes of the analysed papers. Multiword keywords are multiword noun phrases of varying length (2,3,4-word phrases) that are (statistically) typical of a corpus[3]. It was decided not to report one-word keywords as they tend to reflect more proper names in academic discourse (surnames and years) than in other registers (Biber & Gray, 2016).

4. Results

4.1 Number of studies addressing DDL and corpora and scope of interest (RQ 1)

Only 32 of the 759 papers published in the five journals during the 2011–2015 period explored the use of DDL and different types of corpora for language learning. DDL and corpora in language learning and teaching represented 4.2% of the published research. Table 1 shows the total number and percentage of papers examining DDL and corpora for language learning and teaching from 2011 to 2015:

Journal	Total	DDL/corpora in language learning and teaching	%
CALL	133	11	8.3
CALICO Journal	113	6	5.3
LLT	80	2	2.5
ReCALL	88	10	11.4
System	345	3	0.9
Total	**759**	**32**	**4.2**

[3] URL: https://www.sketchengine.eu/wp-content/uploads/ske-statistics.pdf

Table 1. Research published from 2011 to 2015 in top-ranked CALL journals addressing DDL and corpora in language learning and teaching

However, significant differences among journals were observed. *ReCALL* (11.4%) and *CALL* (8.3%) published the highest number papers addressing DDL/corpora in language learning during the aforementioned period, although the former devoted a special issue to DDL in 2014 which included eight papers, thus explaining the peak that year. Two *ReCALL* papers were not retained for further analysis: the first did not research language corpora (Caws, 2013), whereas the second used corpora as a research tool instead of a learning or teaching resource (Farr, 2015). *System* (0.9%), *LLT* (2.5%) and *CALICO Journal* (5.3%) featured the lowest numbers of published research in our area of interest. The percentage for *System* is not surprising given that it is a more generalist journal than the others. Two *LLT* papers actually used corpora and corpus linguistics, but they were not included in our analysis because the focus was on automatic grading of learner language (Crossley et al., 2012) and the analysis of learner error in computer-mediated communication (MacDonald, 2013), respectively. One paper in *CALICO Journal* (Hubbard, 2013) was primarily concerned with learner training in research, development, practice, and teacher education. Although Hubbard (2013: 173) claims that one of the more developed areas of learner training for CALL is "teaching students strategies for utilizing corpora and concordance programs to engage in data-driven learning", this paper was not considered as it did not specifically address the use of DDL or corpora for language learning or teaching.

A multiword keyword analysis (Kilgariff, 2012; Pérez-Paredes, 2017) of all 32 papers revealed the top ten terms in our set of papers to be *corpus use*, *corpus consultation*, *data-driven learning*, *learner corpus*, *academic writing*, *second language*, *language learning*,

second language acquisition, *experimental group* and *collocation retrieval.* The top 50 multiword keywords can be found in Appendix 2. These keywords indicate that there is a clear focus on (a) writing as the main target skill; (b) concordancing; and (c) collocations. Most of the keywords relate to learning, SLA and a procedural focus rather than the challenges of using technology (Chambers & Bax, 2006).

4.2 Examining normalisation factors (RQ 2)

In the following section, we will offer a breakdown of the factors that may impede or facilitate normalisation as conceptualised in Chambers & Bax (2006). We will refer to each paper by ID number to aid readability of what follows.

4.2.1 Logistics

The term *logistics* is rarely discussed or even mentioned in the papers under analysis. In Chambers & Bax (2006), *logistics* refers to resource location and access, room layout and the lack of time to use such resources. In the papers examined, when a computer laboratory was used, we did not find many criticisms or explicit complaints about the limited access to equipment or the designated rooms. Most of the DDL experiments carried out in computer labs did not include specific references to layout or equipment used. Smith (2011: 300) reported that "all students seemed to enjoy the CALL laboratory sessions" over a semester period, and Chen (2011: 65) had informants test and use "different corpus-based tools in a computer laboratory for about three hours". Pérez-Paredes et al. (2011) conducted their experiment in a computer lab over a week across three sessions, whereas Chang (2014: 246) used an engineering lab while "serving as an English writing instructor". Lai & Chen (2015) carried out their study during an EFL introductory writing class in a computer lab two hours per week for 16 weeks. In Cowan et al. (2014), the CALL group received computer

instruction once a week for four weeks in the language lab, and Lénko-Szymanska's (2014) sessions took pace in a lab with internet connection and access to Moodle. Only Daskalovska (2015) suggested that the distribution of her informants was conditioned by computer lab availability. However, we found that the use of some software was not totally exempt from difficulties. For example, Wood (2011) reported that he could not get IT support from his university to set up a server to implement a web application.

The physical space in which corpora are used does not seem to play a crucial – be it impeding or particularly facilitative – role in the use of DDL and corpora for language learning. Instead, what we found were different approaches to the notion of *access*, ranging from the development and testing of new resources to the use of well-known corpus resources. The analysed research seems to range from the belief that new ad-hoc software and corpora (i.e. resources developed by the researchers) need to be developed as a response to meeting students' needs (c.f. Chang, 2014) to the adaptation and (re)use of popular corpora available on the Internet (i.e. Mark Davies´ BNC & COCA distributions) to improve the learning of different skills, vocabulary and grammatical aspects, including learner errors (c.f. Smart, 2014). A considerable number of research papers [IDs 1, 8, 12, 13, 18, 22, 26 and 37] explicitly report the development and testing of new software (i.e. new software developed to by the research team) that attempts to fill the gap in DDL or corpora use for language learning. These papers account for 25% of the research analysed. Six papers used the British National Corpus (BNC) [IDs 2, 6, 11, 12, 32 and36]; five used the Corpus of Contemporary American English (COCA) [IDs 17, 28, 30, 31 and 32]; and six papers compiled their own corpus to be used with language learners [IDs 4, 10, 23, 27, 29 and 35]. Five further papers used other corpora including the Michigan Corpus of Academic Spoken English (MICASE) [ID 32]; the Michigan Corpus of Upper-level Student Papers (MICUSP) [IDs 17 and 36]; the

British Academic Written English Corpus (BAWE) [ID 36]; and Google as a search engine [ID 5]. Figure 2 shows the percentage of different corpus resources used in the research analysed.

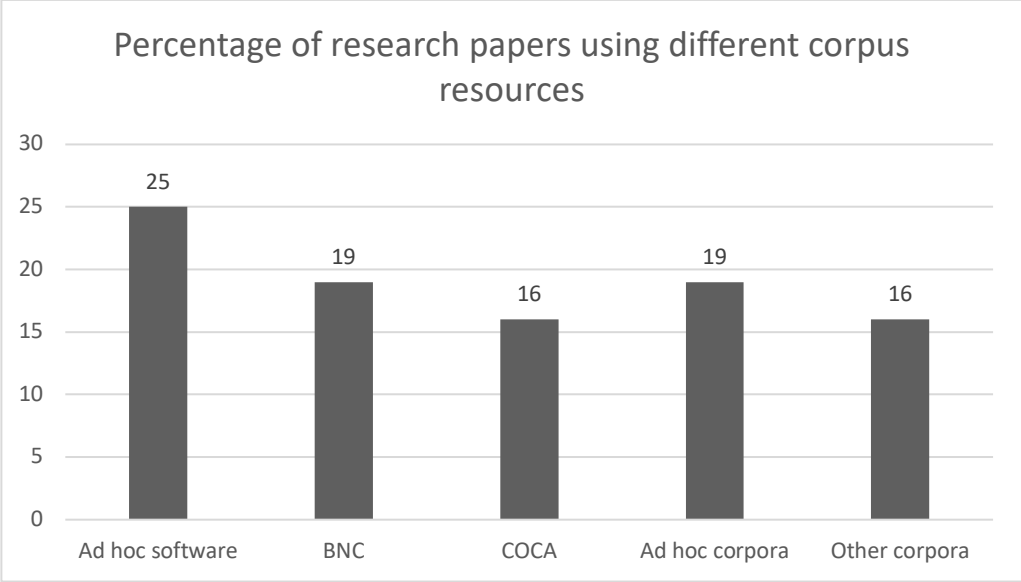

Figure 2. Papers using different corpus resources

In some of the research analysed, the learning activities were either implemented on the Moodle open-source virtual learning environment [IDs 2 and 7] or via the use of a server and ad hoc software [IDs 3 and 13].

4.2.2 Stakeholders' conceptions, knowledge and abilities

Chambers & Bax (2006) examined teachers' conceptions, knowledge and abilities and interpreted them as either impeding or facilitating factors. However, their use of these terms remains vague and, for the most part, lacks a clear theoretical underpinning.[4] We decided to look at attitudes and abilities as they emerge in the papers under analysis, and will use *knowledge focus* to examine the procedural skill-based knowledge needed by students and teachers to query, use and interpret concordance lines and, generally, work with corpora.

[4] Despite mentioning the term, *knowledge* is not discussed in Section 7.2 of Chambers & Bax (2006).

Most of the research examined language learners' attitudes towards the use of DDL and corpus resources for language learning. In 69% of the papers analysed [IDs 1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 18, 21, 27, 29, 30, 31, 32, 34, 35 and 37], corpus and DDL users were invited to express their attitudes towards corpus resource use for language learning, largely through questionnaires but also through interviews. The vast majority of students taking part in these studies found that DDL and corpora use was useful for their learning of vocabulary and collocational behaviour [IDs 6, 7, 8, 11, 12, 14, 27 and 37]; their writing [IDs 10, 29 and 31]; their speaking [ID 30]; and their register awareness [IDs 2 and 35]. Some research examined structure recognition and morphology, yielding similar positive reactions [IDs 13 and 34]. The idea of "usefulness" seems to be central to the analysis of students' conceptions of corpora use. Despite the generally positive reactions reported, we need to show caution in our use and interpretation of methods that capture students' opinions using closed-ended questionnaires. In papers such as Aguado et al. (2012), the use of Likert scales is biased towards obtaining positive results, given that the statements mainly reflect the benefits of using DDL and none of the associated challenges or difficulties. Smith (2011: 307) reported that "seven of the 19 projects made a negative comment of some sort", with tediousness and lack of understanding of how to compare frequencies across corpora singled out as students' grounds for concern. Wood (2011: 672) also highlighted criticism from one of the participants in the study, namely a "retired professor in Humanities [who] had an intense dislike for computers in general and became frustrated quickly with the program". Some of the language learners' perceptions echo the challenges of interpreting concordance lines. Geluso & Yamaguchi (2014) reported that students found the use of COCA and, particularly, the interpretation of cut-off sentences extremely demanding.

The vast majority of learners participating in these studies were university students. In 94% of the research papers, HE informants were used across different research designs. Only two papers (IDs 25 and 34) explored the use of DDL and corpora among secondary school learners. None of the papers examined the role of management across the institutions where the experiences were carried out, which suggests that management is not perceived as an "obstacle to successful normalisation" (Chambers & Bax, 2006: 473).

While *ability* in Chambers & Bax (2006) denotes *computer* competence, in the papers analysed the notion of ability refers to either skill ability or specific abilities when using DDL or corpora. An exploratory collocational analysis of the lemma *ability* in the papers revealed the following learner competencies as relevant to the discussion:

- to consult an online learner dictionary quickly and efficiently;
- to make generalisations about usage;
- to edit grammatical errors from (learners') writing; and
- to use L2 collocations.

These abilities are generally presented as facilitating language learning and are central to the learning tasks that students are expected to accomplish. *Cognitive abilities* are discussed either in the literature review or in the discussion sections of the papers, but they are not usually the main focus of research, that is, they are not part of the research questions addressed. The term *cognitive skills* is used by Yoon & Jo (2014) and in other papers when discussing O'Sullivan's (2007) list of corpus-related skills.[5]

[5] According to O'Sullivan (2007: 277), these skills are predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying.

In terms of *knowledge*, the learners in these papers are expected to understand and ultimately acquire a wide range of both procedural and declarative knowledge-related skills. Table 2 summarises the *knowledge focus* across all 32 papers.

Paper id	Knowledge focus
1	Collocations + search POS frequency
2	Corpus use + search and frequency
3	Parallel concordance lines
4	Compilation of a corpus + mixed analytical abilities
5	Search Frequency
6	Concordancing + exploring and noticing
7	Clause patterns + search and structure recognition
8	Concordancing + collocation
9	Dictionary skills + search skills + collocations
10	Writing + Paraphrasing
11	Collocations and concordance skills
12	Writing and form awareness
13	Reading + morphology + collocations
14	Vocabulary acquisition + Usage
15	Recognising patterns
17	Productive knowledge of linking adverbials + writing + collocational competence and text register awareness
18	Idioms + collocations
21	Grammatical and lexical accuracy
22	Grammatical and lexical accuracy + error recognition
23	Developing authorial stance in advanced academic writing
25	Using dictionaries

26	Writing + correction of grammatical and lexical error after feedback at revision stage
27	Writing + lexico grammatical use of abstract nouns
28	Corpus-informed grammar + passive and error correction tasks
29	Writing and use of linking adverbials
30	Formulaic language + pattern hunting
31	Corpus-aided writing
32	Using corpora for language teaching
34	Grammatical analysis of POS
35	Register awareness + frequency analysis + Linguistic features analysis
36	Academic literacy + register awareness
37	Metalinguistic awareness + detecting errors

Table 2. Knowledge focus of the papers analysed

Most of the research used DDL and corpora to improve students' writing by examining collocations and language patterning via a range of procedural knowledge as described in Table 2. The learners' analysis and evaluation of frequency remained an important knowledge item and few papers studied specific grammatical constructions (i.e., passive and abstract nouns).

4.2.3 Syllabus integration

Chambers and Bax (2006: 478) suggest that "successful normalisation of CALL requires that it be properly integrated into the syllabus". Only a handful of studies [IDs 9, 30, 32, 34 and 35] discussed the integration of DDL and corpora across syllabi. Lénko-Szymanska (2014: 263) developed a syllabus for the course "Corpora in Foreign Language Teaching" offered to MA students at the University of Warsaw and was designed to introduce "the concept of a corpus and its analysis, and to outline various applications of corpora in language education".

Geluso (2014) developed a syllabus where students were introduced to DDL and formulaic language. However, two of the papers that explicitly address the integration of DDL and corpora in language education do so to suggest that syllabus integration would be beneficial for language learners. Aguado et al.(2012) argue that the use of spoken learner and native speaker corpora can help students achieve a more natural oral production, and Lin (2015) recommends that language teachers use spoken features extracted from corpora in their teaching. In Lai & Chen (2015), the *syllabus* is the regular, non-DDL syllabus whereby researchers try to integrate CALL by using different online corpus and dictionary tools and websites. References to *integration* are vague and usually reflected in the literature review or in the discussion sections rather than in the methodology paragraphs (e.g., Geluso, 2014; Ranalli, 2013). While Comelles et al. (2013) claim that integrating corpus applications in the language classroom facilitates reflection on genuine data, Pérez-Paredes et al. (2011) suggest that DDL can benefit from its integration with online resources whose use is more normalised in language education, mainly search websites and dictionaries. Other researchers, however, argue that their experiments confirm successful integration of DDL and corpora. Gordani (2013: 441) used an online corpus-based approach integrated into 42 hours of reading comprehension classroom instruction. The author claims that "the main effect of corpus integration has been significant", given that the experimental DDL group obtained better results at post-test.

4.2.4 Training and support

Chambers & Bax (2006) found that language teachers at their two research sites needed further training and development when using CALL. In particular, they suggested that collaborative rather than expert-to-novice training would be beneficial. Our analysis shows that *training* is an important theme; 60% of papers [IDs 1, 2, 3, 4, 5, 6, 9, 11, 12, 14, 15, 18,

21, 26, 30, 31, 32, 34 and 37] included some form of discussion in this area. Most papers incorporated references in their respective literature reviews (Chen, 2011; Gao, 2011; Geluso, 2013; Gordani, 2013; Ranalli, 2013; Deluso & Yamaguchi, 2014; Tono et al., 2014) to lend support to the need for learner training in corpora use for language learning. All of them, albeit differently, provided training mainly to students but also to teachers (Lénko-Szymanska, 2014; Yoon & Jo, 2014). Students' training ranged from minimal training in Daskalovska (2015) to intensive sessions (Amer, 2014; Vhang, 2014) or deliberate no training (Poole, 2012). In Pérez-Paredes et al. (2011), learner training was subsumed under the notion of corpus consultation guidance.

Support is a prerequisite for normalisation as teachers' lack of skills and technical failures of CALL-related equipment need to be addressed by institutions (Chambers & Bax, 2006). In our review, 28% of the analysed papers discuss *support* [IDs 8, 10, 13, 15, 23, 26, 28, 30, 36 and 37]. In some of them, the corpus itself provides support to learners when writing (Chen et al., 2015) or correcting errors (Tono et al., 2014). In Rezaee et al. (2015), support is understood as scaffolding between peers and teachers, whereas Tribble & Wingate (2013) situated their research in the context of literacy support provided by HE institutions in the UK. Some studies suggested that students needed further support (Chang, 2012) to infer patterns. Meanwhile, Wood (2012) was the only study that mentioned insufficient technical support at one of the research sites.

5. Discussion

Our research used the normalisation factors in Chambers & Bax (2006) as the framework for the systematic review of 32 papers investigating DDL and corpora in CALL-related research journals during the 2011–2015 period. Out of the 759 research papers published in the five CALL journals analysed during this period, only 4.2% of papers examined corpora or DDL in language education. Distribution varied from journal to journal, ranging from 0.9% in *System* to 11.4% in *ReCALL*. In terms of the papers' main research focus, this was largely, yet not exclusively, found to be DDL-assisted writing. Although most terms extracted from these papers are more concerned with learning and pedagogy than with computer technology, we observed the main focuses as the affordances of the tools (i.e., concordance lines) and corpora (i.e., language patterning emerging from the variety of corpora used by researchers), their use and their impact on mostly short-term language gains and corpus analysis skills.

In the papers analysed, we found that logistics plays a very different role to the original discussion in Chambers & Bax (2006). Access to physical space, or the use of language labs, is not perceived as a significant or impeding factor in any of the papers we surveyed. When labs are used, there is never a sense that either the students or researchers themselves are experiencing technology-related problems as a result of equipment or software use. This might suggest that our researchers are experts in the use of both DDL and computers in language education. Furthermore, this idea is supported by the fact that researcher training is not discussed in the research under analysis. We argue, however, that these researchers understand *access* as the provision of relevant corpora to language learners. We found that 25% of the research analysed introduces new tools or new software developed for use in the language classroom. In terms of the corpora used, 19% of research papers make use of new corpus resources developed for specific groups of language learners. These percentages show

that there is a very high degree of innovation and agency among the researchers surveyed. Because faculty or school management is not explicitly seen as a relevant factor in using DDL, we would deem the main stakeholder in DDL use to be researchers rather than institutions or even language teachers. This finding is supported by the fact that 94% of the surveyed research was performed at universities where researchers presumably have easier access to samples. These results confirm trends in previous research: Boulton (2008) and Boulton & Cobb (2017) showed that ony a small percentage of research into DDL involved non-university students, whereas Pérez-Paredes (2010) highlighted the sample bias of HE humanities students in DDL research, which may raise validity issues.

In terms of the stakeholders' conceptions, 69% of the research reported data that examine the uses of corpora in the language classroom. Most learners seem to endorse the usefulness of DDL for vocabulary and collocations, although students encounter different challenges in using the software, especially when interpreting the concordance lines, which is consistent with findings in previous research (c.f. Boulton & Pérez-Paredes (2014) and Vyatkina & Boulton's (2017) introductions to the special issues in *ReCALL* and LLT, respectively). While researchers largely focused on language gains and DDL effectiveness, research examining the cognitive abilities associated with the use of concordancers was not represented in the body of research examined. As for language syllabi, only 16% of the research papers addressed syllabus integration. When integration is mentioned, it is often within the context of experiments rather than within the larger context of the language learners' curriculum. In terms of training and support, 60% of the research papers either echoes the concerns in the literature about the need to train learners in language corpora use or discusses the importance of training students to perform during the experiment. Most papers that mention the role of support do so to highlight corpora as providers of support for

language learners: training is conceptualised as a process that enables learners to operate the software and understand the concordance format.

Our results suggest that this body of research tends to present DDL and corpora as a solution to some of the problems or shortcomings in language education. This representation of DDL emerging from the papers echoes the concerns voiced by Bax (2003) about CALL research which tends to showcase software and technology as providing packaged solutions to language learning and access to linguistic knowledge as a semiotic resource, while minimising the micro levels of language learning where cognitive capacities are developed (Douglas Fir Group, 2016). The following quotes from two of the papers under study place focus on the tool and DDL *technology* rather than on the meaning potential of the semiotic resources (Douglas Fir Group, 2016) afforded by DDL:

> This study has demonstrated that advanced learners can easily learn how to use concordance software and obtain the desired information successfully even with minimal training. (Daskalovska, 2015)

> At the same time concern ranged among the students from the workload imposed
> to the anxiety in dealing with technology and large amounts of data. (Gordani, 2013)

Whereas Chambers & Bax (2006) focused on technology-related factors, our body of research suggests that DDL researchers favour a tool-oriented discourse that revolves around the affordances of corpora and learners' perceived usefulness of said corpora. However, this is at the expense of theorisation on the role of corpora and DDL in second language learning. The field of DDL, as represented in our body of research, offers a rather complex picture where, while some of the obstacles and impeding factors have been removed (technical issues, access to resources, teachers' training), there is still a dearth of debate around (1) the

role of DDL in language learning theory and (2) the roles of non-researcher language teachers as facilitators of CALL integration (Martins & Moreira, 2016). Future research should address how DDL can be accommodated within different second language learning theories (Flowerdew, 2015), in particular usage-based approaches to language learning. A focus on theorisation may contribute to our understanding of how DDL may strengthen learners' "new symbolic constructions" (Ellis, Römer & O'Donnell, 2016: 23) and promote implicit learning and automatisation of the language system. Most research designs in our review have focused on the perceived usefulness of DDL applications. However, further research should shed some light on how, among other things, the role of the frequency of constructions as shown in DDL and learners' interaction with different types of constructions impacts language learning beyond post-delayed tests.

Bax (2011) offered a reformulation of the normalisation notion in the context of neo-Vygotskyan sociocultural theory principles, mainly derived from Mercer & Fisher (1997), and discussed how effective educational practice can benefit from a set of elements including not only access, participation and interaction with sources and other learners, but also expert intervention, which scaffolds, models and challenges learning. These so-called elements can thus be seen as playing a mediational role in CALL and offer "a foundation for the more practical domain of planning for normalisation" (p.11). While sociocultural theory has not been explored extensively in DDL studies, with the exception of Rezaee et al. (2015), the potential role of DDL in usage-based accounts of language learning, while promising, remains largely unexplored. Boulton & Cobb (2017) highlighted the implications of using DDL for pattern-based learning and chunking (pp. 350-351). Our analysis of the multiword keywords used by researchers has brought about interest in usage-based topics such as collocations and formulaic language which could be better understood if they were

theoretically framed. However, these and other theorisation avenues were rarely explored in the research under analysis. The fact that the pool of papers examined is fundamentally empirical may have contributed to a lack of "theoretical positioning" (Hanks, 2019: 143) that we may find in other DDL research outside the scope of the journals analysed (i.e. other journals, books or book chapters). Geluso (2013), Geluso & Yamaguchi (2014) and Huang (2014) are the only exceptions. We argue that such a lack of theorisation may keep DDL research in a loop where usefulness and language gains are targeted as main research questions, which prevents mainstream language teachers from understanding DDL practice in the wider contexts of SLA and language education. Possible constructs that could be addressed within usage-based theories may include, among others, the role of frequency and dispersion in the DDL input, categorisation and prototype effects of such input, or research that addresses the role of explicit knowledge in language development as usage based views hold that "the bulk of language learning happens implicitly" (Tyler & Ortega, 2018: 318).

While the learner's voice is present in some of the analysed papers, the role of language teachers is subsumed by researchers, which prevents us from gaining a better grasp of how DDL can be normalised in contexts where language teachers work for non-HE institutions or they do not wish to carry out semi-experimental or mixed methods research. Arguably, action research or exploratory practice (Hanks, 2019) may increase the visibility of DDL across instructional contexts, languages and levels. As Warren (2016: 343) put it: "The difficulties encountered […] do not necessarily negate the DDL approach to language learning but rather underline the need for larger and more comprehensive corpora in order to better support corpus linguistics and data-driven learning". Using normalisation as a framework demonstates how advances in DDL can focus more on the L (learning) and less on the D

(data) while seeking to understand how language teachers and learners can shift from a driven (D), passive perspective to a more agentive and active, learning-centered DDL.

6. Conclusions

One of the advantages of our systematic review lies in the fact that it presents a methodology that could be replicated by other researchers with other journals or forums and/or in other periods (i.e. 2016-2021). We used normalisation as a framework to analyse research into DDL and corpus use in language learning and teaching in five CALL-related journals during the 2011–2015 period. Based on our analysis, we can conclude that DDL normalisation in language education has only taken place in a limited number of contexts where language teachers and DDL researchers subsume the same roles in HE, particularly in Asia, Europe and the US. The body of research under analysis tends to favour quantitative research methods. Within this scope, we have identified two areas in which DDL is far from normalised: syllabus integration and language teacher training. These areas are rarely discussed in the papers under review and, based on this evidence, we argue that a lack of normalisation may be linked to the fact that research into DDL and corpora use in language learning is, according to Vyatkina & Boulton (2017: 2) "still developing".

While the findings of this systematic review may contribute to advancing our understanding of DDL in general, and of DDL integration (Martins & Moreira, 2016) in particular, there are some limitations with this study. First, the research encompasses only published papers in five journals, which excludes research in other journals or other formats during the same time period. All systematic reviews follow a set of criteria and, in this paper, research not published in the five journals that were surveyed has not been considered. Our findings

should be, therefore, confined to the scope of the research published in these journals during the 2011-2015 period. Other DDL research within the same time line and published elsewhere will not necessarily reflect the range of concerns addressed in this paper. Second, and as with any systematic review, our research paper analysis was based on our coding which, by definition, is biased. The use of multiword keyword analysis intended to reduce this research bias. Despite these limitations, our paper offers a robust picture of the research carried out on DDL and corpora in language learning and teaching in the first half of the 2010s. A study of other periods, or the replication of this research by other researchers, can only enhance our understanding of "language learning and teaching in our increasingly networked, technologized and mobile worlds" (Douglas Fir Group, 2016: p. 20).

Acknowledgement

I would like to thank the anonymous reviewers for their insightful and constructive comments on this paper. Special thanks to Lindsay Duffy, Geraldine Mark and Anne O´Keeffe for their suggestions.

Notes on contributors

Pascual Pérez-Paredes is a lecturer at the Faculty of Education, University of Cambridge. His research interests include CALL, learner language variation, corpora in language education and corpus-assisted discourse analysis. He is currently an editorial board member of ReCALL and Register Studies. He has published in journals such as ReCALL, CALL, Language, Learning & Technology, System, Journal of Pragmatics and English for Specific Purposes.

ORCID

Pascual Pérez-Paredes: https://orcid.org/0000-0002-2796-338X

References

Aguado-Jiménez, P., Pérez-Paredes, P., & Sánchez, P. 2012. Exploring the use of multidimensional analysis of learner language to promote register awareness. *System*, 40(1), 90-103.

Amer, M. 2014. Language learners' usage of a mobile learning application for learning idioms and collocations. *CALICO Journal*, 313, 285-302.

Bax, S. 2003. CALL—past, present and future. *System*, 311, 13-28.

Bax, S. 2011. Digital education: Beyond the "wow" factor. In *Digital Education.* Palgrave Macmillan, New York, pp. 239-256.

Braun, S. 2005. From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 171, 47-64.

Boulton, A. 2008. "Evaluating corpus use in language learning: state of play and future directions". Paper presented at the American Association of Corpus Linguistics. Provo, March 13-15, 2008.

Boulton, A. & Cobb, T. 2017. Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348-393.

Boulton, A. & Pérez-Paredes, P. 2014. ReCALL special issue: Researching uses of corpora for language teaching and learning Editorial Researching uses of corpora for language teaching and learning. *ReCALL*, 26(2), 121-127.

Caws, C. G. 2013. Evaluating a web-based video corpus through an analysis of user interactions. *ReCALL*, 251, 85-104.

Chambers, A., & Bax, S. 2006. Making CALL work: Towards normalisation. *System*, 344, 465-479.

Chang, J. Y. 2014. The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, 262, 243-259.

Chang, P. 2012. Using a stance corpus to learn about effective authorial stance-taking: a textlinguistic approach. *ReCALL*, 242, 209-236.

Chen, H. J. H. 2011. Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, 241, 59-76.

Chen, M. H., Huang, S. T., Chang, J. S., & Liou, H. C. 2015. Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer Assisted Language Learning*, 281, 22-40.

Comelles, E., Laso, N. J., Forcadell, M., Castaño, E., Feijóo, S., & Verdaguer, I. 2013. Using online databases in the linguistics classroom: dealing with clause patterns. *Computer assisted language learning,* 263, 282-294.

Cotos, E. 2011. Potential of automated writing evaluation feedback. *Calico Journal,* 282, 420-459.

Cotos, E. 2014. Enhancing writing pedagogy with learner corpus data. *ReCALL*, 262, 202-224.

Cowan, R., Choo, J., & Lee, G. S. 2014. ICALL for improving Korean L2 writers' ability to edit grammatical errors. *Language, learning & technology,* 183,193-207.

Crossley, S., & McNamara, D. 2013. Applications of text analysis tools for spoken response grading. *Language, learning & technology*, 172, 171–192.

Daskalovska, N. 2015. Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 282, 130-144.

Davies, G., Otto, S. E., & Rüschoff, B. 2014. Historical perspectives on CALL. In Thomas, M., Reinders, H., & Warschauer, M. Eds..Contemporary *Computer-assisted language learning.* London: Bloomsbury., pp. 19-38.

Douglas Fir Group Atkinson, D.; Byrnes, H.; Doran, M.; Duff, P.; Ellis, Nick C.; Hall, J. K.; Johnson, K.; Lantolf, J.; Larsen-Freeman, D.; Negueruela, E.; Norton, B.; Ortega, L.;

Schumann, J.; Swain, M.; Tarone, E. 2016. A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal*,100, 19-47.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). Constructions and usage-based approaches to language acquisition. *Language Learning*, 66, 23-44.

Farr, F., & Riordan, E. 2015. Tracing the reflective practices of student teachers in online modes. *ReCALL*, 271, 104-123.

Frankenberg-Garcia, A. 2014. The use of corpus examples for language comprehension and production. *ReCALL*, 262, 128-146.

Gao, Z. M. 2011. Exploring the effects and use of a Chinese–English parallel concordancer. *Computer Assisted Language Learning*, 243, 255-275.

Garner, J. R. 2013. The use of linking adverbials in academic essays by non-native writers: How data-driven learning can help. *CALICO Journal*, 303, 410.

Geluso, J. 2013. Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing. *Computer Assisted Language Learning*, 262, 144-157.

Geluso, J., & Yamaguchi, A. 2014. Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, 262, 225-242.

Gimeno-Sanz, A. 2016. Moving a Step Further from "Integrative CALL". What's to Come? *Computer Assisted Language Learning*, 29(6), 1102-1115.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. 2014. Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer assisted language learning*, 27(1), 70-105.

Gordani, Y. 2013. The effect of the integration of corpora in reading comprehension classrooms on English as a Foreign Language learners' vocabulary development. *Computer Assisted Language Learning*, 265, 430-445.

Gough, D., Oliver, S., & Thomas, J. 2012. *An introduction to systematic reviews.* London: Sage.

Grgurović, M., Chapelle, C. A., & Shelley, M. C. 2013. A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, 25(2), 165-198.

Hanks, J. 2019. From research-as-practice to exploratory practice-as-research in language teaching and beyond. *Language Teaching*, 52(2), 143-187.

Huang, L. 2011. Corpus-aided language learning, *ELT Journal*, 65,(4), 481–484.

Huang, Z. 2014. The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL*, 262, 163-183.

Hubbard, P. 2013. Making a case for learner training in technology enhanced language learning environments. *Calico Journal*, 302, 163-178.

Johns, T. 1990. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, 10, 14–34.

Kilgarriff, A. 2012. Getting to know your corpus. In International conference on text, speech and dialogue. Berlin: Springer, pp. 3-15.

Lai, S. L., & Chen, H. J. H. 2015. Dictionaries vs concordancers: actual practice of the two different tools in EFL writing. *Computer Assisted Language Learning,* 284, 341-363.

Lee, H., Warschauer, M., & Lee, J. H. 2018. The Effects of Corpus Use on Second Language Vocabulary Learning: A Multilevel Meta-analysis. *Applied Linguistics*. DOI: https://doi.org/10.1093/applin/amy012

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*, 262, 260-278.

Lin, Y. L. 2015. Using key part-of-speech analysis to examine spoken discourse by Taiwanese EFL learners. *ReCALL*, 273, 304-320.

Luo, Q. 2016. The effects of data-driven learning activities on EFL learners' writing development. *SpringerPlus*, 51, 1255.

MacDonald, P., Garcia-Carbonell, A., Carot & Sierra, J. M. 2013. Computer learner corpora: analysing interlanguage errors in synchronous and asynchronous communication. *Language, learning & technology*, 172, 36-56.

Martins, C. B. M., & Moreira, H. 2015. Factors that determine CALL integration into Modern Languages Courses in Brazil. In Sanz, A. M. G., Levy, M., Blin, F., & Barr, D. Eds. WorldCALL: *Sustainability and computer-assisted language learning*. London: Bloomsbury, pp.103-118.

Mizumoto, A., & Chujo, K. 2015. A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, 22, 1-18.

Pérez-Paredes, P. 2010. Corpus Linguistics and Language Education in Perspective: Appropriation and the Possibilities Scenario. In T. Harris & M. Moreno Jaén (Eds.)*, Corpus Linguistics in Language Teaching.* Bern: Peter Lang, pp. 53-73.

Pérez-Paredes, P. 2017. A Keyword Analysis of the 2015 UK Higher Education Green Paper and the Twitter Debate. In M. A. Orts, Breeze, R., & Gotti, M. (Eds.). *Power, persuasion and manipulation in specialised genres: providing keys to the rhetoric of professional communities*. Bern: Peter Lang, pp. 161-193.

Pérez-Paredes, P., Sánchez-Tornel, M., Alcaraz Calero, J. M., & Jiménez, P. A. 2011. Tracking learners' actual uses of corpora: guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, 24,3, 233-253.

Pérez-Paredes, P., Ordoñana, C. & Aguado, P. 2018. Language teachers' perceptions on the use of OER language processing technologies in MALL. *Computer Assisted Language Learning* 31,5-6,522-545.

Poole, R. 2012. Concordance-based glosses for academic vocabulary acquisition. *CALICO Journal,* 294, 679.

Ranalli, J. 2013. Designing online strategy instruction for integrated vocabulary depth of knowledge and web-based dictionary skills. *CALICO Journal*, 301, 16-43.

Rezaee, A. A., Marefat, H., & Saeedakhtar, A. 2015. Symmetrical and asymmetrical scaffolding of L2 collocations in the context of concordancing. *Computer Assisted Language Learning*, 286, 532-549.

Römer, U. 2006. Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik*, 542, 121-134.

Sinclair, J. 2003 *Reading Concordances.* London: Longman.

Smart, J. 2014. The role of guided induction in paper-based data-driven learning. *ReCALL*, 262, 184-201.

Smith, S. 2011. Learner construction of corpora for general English in Taiwan. *Computer Assisted Language Learning,* 244, 291-316.

Steel, C. H., & Levy, M. 2013. Language students and their technologies: Charting the evolution 2006–2011. *ReCALL*, 25(3), 306-320.

Sun, B., & Ye, C.C. 2006. Reassessment of computer-assisted language learning. *Computer Assisted Foreign Language Education*, 110, 98–110.

Thomas, M., Reinders, H., & Warschauer, M. (Eds.). 2014. *Contemporary computer-assisted language learning.* London: Bloomsbury.

Tono, Y., Satake, Y., & Miura, A. 2014. The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, 262, 147-162.

Traxler, J. & Kukulska-Hulme, A. 2016. *Mobile Learning: The Next Generation.* London Routledge.

Tribble, C. 2008. From Corpus to Classroom: Language Use and Language Teaching. *ELT Journal,* 62, 213-216.

Tribble, C., & Wingate, U. 2013. From text to corpus–A genre-based approach to academic literacy instruction. *System*, 412, 307-321.

Vyatkina, N., & Boulton, A. 2017. Corpora in language learning and teaching. *Language Learning & Technology*, 213, 1–8.

Warschauer, M., 2000. CALL for the 21st Century. IATEFL and ESADE Conference, 2 July 2000, Barcelona, Spain.

Warren, M. 2016. Introduction to data-driven learning. In Farr, F. & Murray, L. (Eds.) *The Routledge Handbook of Language Learning and Technology*. London: Routledge, pp. 337-348.

Wood, P. 2011. Computer assisted reading in German as a foreign language, developing and testing an NLP-based application. *CALICO Journal,* 283, 662-676.

Yang, Y. F., Wong, W. K., & Yeh, H. C. 2013. Learning to construct English L2 sentences in a bilingual corpus-based system. *System*, 413, 677-690.

Yoon, H., & Jo, J. 2014. Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology*, 181,96-117.

Appendix 1. Research papers analysed

Authors	id	Journal	Year	Title
Chen, H. H.	1	CALL	2011	Developing and evaluating a web-based collocations retrieval tool for EFL students and teachers
Pérez-Paredes et al.	2	CALL	2011	Tracking learners´ actual uses of corpora: guided vs non-guided corpus consultation
Gao, Z. M.	3	CALL	2011	Exploring the effects and use of a Chinese-English parallel concordancer
Smith, S.	4	CALL	2011	Learner construction of corpora for general English in Taiwan
Geluso, J.	5	CALL	2013	Phraseology and frequency of occurrence on the web: native speakers´ perceptions of Google-informed second language writing
Gordani, Y.	6	CALL	2013	The effect of the integration of corpora in reading comprehension classrooms on English as a Foreign Language learners´ vocabulary development
Comelles et al.	7	CALL	2013	Using online databases in the linguistics classroom: dealing with clause patterns
Rezaee et al.	8	CALL	2015	Symmetrical and asymmetrical scaffolding of L2 collocations in the context of concordancing
Lai, S. & Chen, H. H.	9	CALL	2015	Dictionaries vs concordancer: actual practice of the two different tools in EFL writing
Chen et al.	10	CALL	2015	Developing a corpus-based paraphrase tool to improve EFL learners' writing skills
Daskalovska, N.	11	CALL	2015	Corpus-based versus traditional learning of collocations
Cotos, E.	12	CALICO	2011	Potential of automated writing evaluation feedback
Wood, P.	13	CALICO	2011	Computer assisted reading in German as a Foreign Language. Developing and testing an NLP-based application
Poole, R.	14	CALICO	2012	Concordance-based glosses for academic vocabulary acquisition
Ranalli, J.	15	CALICO	2013	Designing online strategy instruction for integrated vocabulary depth of knowledge and web-based dictionary skills
Hubbard, P.	16	CALICO	2013	Making the case for learner training in technology enhanced language learning environments
Garner, J. R.	17	CALICO	2013	The use of linking adverbials in academic essays by non-native writers: how data-driven learning can help
Amer, M.	18	CALICO	2014	Language learners´ usage of a mobile learning application for learning idioms and collocations
Crossley, S., & McNamara, D	19	LL&T	2013	Applications of text analysis tools for spoken response grading
MacDonald et al.	20	LL&T	2013	Computer learner corpora: analysing interlanguage errors in synchronous and asynchronous communication
Yoon & Jo	21	LL&T	2014	Direct and indiect ac to corpora: an exploratory case study comparing students´ error correction and learning strategy use in L2 writing

Cowan et al.	22	LL&T	2014	ICALL for improving Korean L2 writers´ ability to edit grammatical errors
Chang, P.	23	ReCALL	2012	Using a stance corpus to learn about effective authorial stance-taking: a textlinguistic approach
Caws, C.G.	24	ReCALL	2013	Evaluating a web-based video corpus through an analysis of user interactions
Frankenberg-García, a.	25	ReCALL	2014	The use of corpus examples for language comprehension and production
Tono et al.	26	ReCALL	2014	The effects of using corpora on revision tasks in L2 writing with coded error feedback
Huang, Z.	27	ReCALL	2014	The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing
Smart, J.	28	ReCALL	2014	The role of guided induction in paper-based data driven learning
Cotos, E.	29	ReCALL	2014	Enhancing writing pedagogy with learner corpus data
Geluso, J., & Yamaguchi, A	30	ReCALL	2014	Discovering formulaic language through data-driven learning: Student attitudes and efficacy
Chang, J. Y.	31	ReCALL	2014	The use of general and specialized corpora as reference sources for academic English writing: A case study
Lénko-Szymanska, A.	32	ReCALL	2014	Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education.
Farr, F., & Riordan, E.	33	ReCALL	2015	Tracing the reflective practices of student teachers in online modes
Lin, Y.L.	34	ReCALL	2015	Using key part-of-speech analysis to examine spoken discourse by Taiwanese EFL learners
Aguado-Jimênez et al.	35	System	2012	Exploring the use of multidimensional analysis of learner language to promote register awareness
Tribble, C., & Wingate, U.	36	System	2013	From text to corpus - A genre-based approach to academic literacy instruction
Yang, et al.	37	System	2013	Learning to construct English (L2) sentences in a bilingual corpus-based system

Shaded references were excluded from the analysis.

Appendix 2. Multiword keywords emerging form the body of research papers analysed

	Keyword	Keyness score	Freq. in the articles analysed	Freq. in the enTenTen13 corpus
1	corpus use	475.31	174	1
2	corpus consultation	393.93	143	0
3	data-driven learning	286.77	104	0
4	learner corpus	273.03	99	0
5	academic writing	230.21	99	21
6	second language	207	205	194
7	language learning	200.46	181	167
8	second language acquisition	159.29	71	26
9	experimental group	147.14	67	29
10	collocation retrieval	141.14	51	0
11	language acquisition	133.09	89	95
12	bilingual concordancer	124.65	45	0
13	language teaching	123.69	148	258
14	learner language	119.15	43	0
15	sentence construction	118.95	46	8
16	corpus analysis	114.13	43	5
17	metalinguistic awareness	101.85	37	1
18	discourse form	99.92	36	0
19	local learner	99.92	36	0
20	retrieval tool	97.5	37	6
21	direct corpus	97.17	35	0
22	collocation retrieval tool	94.42	34	0
23	reading comprehension	93.79	38	14
24	specialized corpus	88.93	32	0
25	writing instruction	88.93	32	0
26	language education	88.91	34	7

27	language proficiency	87.43	34	9
28	language classroom	86.45	35	14
29	local learner corpus	86.18	31	0
30	formulaic language	85.5	31	1
31	parallel concordancer	83.43	30	0
32	vocabulary learning	82.77	30	1
33	vocabulary acquisition	82.04	30	2
34	negative evidence	81.07	32	11
35	control group	80.78	96	256
36	passive voice	80.11	35	24
37	online dictionary	80.04	29	1
38	indirect corpus	77.94	28	0
39	call group	77.94	28	0
40	computer-assisted language	77.94	28	0
41	foreign language	77.2	116	353
42	academic vocabulary	75.19	27	0
43	language pedagogy	72.95	29	12
44	corpus tool	72.44	26	0
45	writing process	71.78	29	14
46	English writing	70.21	27	8
47	stance corpus	69.69	25	0
48	indirect use	69.14	25	1
49	delayed post-test	66.95	24	0
50	indirect corpus use	66.95	24	0