TRABAJO FIN DE GRADO

GRADO EN ESTUDIOS INGLESES: LENGUA, LITERATURA Y CULTURA

[DATA-DRIVEN LEARNING IN THE CONTEXT OF THE EFL CLASSROOM: A SYSTEMATIC REVIEW FOR THE PERIOD 2016-2019]

[AIDA GARCÍA VICUÑA]

[agarcia384@alumno.uned.es]

TUTOR ACADÉMICO: [Elena Martín Monje]

LÍNEA DE TFG: [Aplicaciones de las TIC en los estudios ingleses]

FACULTAD DE FILOLOGÍA

CURSO ACADÉMICO: 2019-20- Convocatoria: [Septiembre]

Abstract

This paper discusses the trends in Data-driven learning (DDL) within the English as a Foreign Language (EFL) context. A total of 26 research papers published during the 2016-2019 period and focused on the matter of DDL were explored in depth. Following Pérez-Paredes' previous review (2019), a corpus-based analysis was also conducted on the selected literature. Results show that the trend on lack of theorisation continues, that research on DDL is scarce and that the majority of it is experimental and effects-oriented. It was also found that research was conducted mainly in East-Asian countries in the context of higher education.

Keywords: DDL, corpora, EFL, second language learning, systematic review

Table of Contents

1. Introduction

The aim of this paper is to review academic literature related to Corpus Linguistics (CL) and its applications in English as a Foreign Language (EFL) teaching, specifically selected bibliography of recent scholarship in Data-Driven Learning (DDL) or 'the learner as a researcher' approach.

In line with some of the works consulted on the subject (McEnery & Xiao, 2011), what follows is a preliminary classification of corpus-based approaches to English language teaching: a) second language acquisition theories and learner corpora, b) indirect use of corpora: designing teaching materials and syllabus, dictionaries and reference grammars, c) direct use of corpora or DDL.

With regard to its current relevance to English Studies, the teaching of English is featured in UNED's Bachelor's Degree guide as one of the main career opportunities for undergraduates. Secondly, CL has a significant role in EFL teaching. As stated by Aijmer (2009), CL researchers "are generally enthusiastic about what they have to offer the teaching profession" (p. 2). Thirdly, DDL is an important field to consider, since some of its affordances include the developing of "a more autonomous learning style" (Guilquin & Granger, 2010, p. 365) as well as lexico-grammatical patterns awareness (Huang, 2014) and it also helps to acquire the knowledge of a language that a native speaker unconsciously possesses, by exploring how words work in context.

1.1. Background

The so-called father of DDL, Tim Johns (1990), drew attention to a newly discovery-based approach where the teacher provides the "context in which the learner is to discover the foreign language" (p. 1). This shift in the traditional focus of learning from the teacher to the learners themselves is directly related to student-centred approaches in second language acquisition theories. Accordingly, DDL is commonly described as a student-centred method focused on presenting to the learner natural and authentic instances of language produced by native speakers (Talai & Fotovatnia, 2012). All in all, teachers are facilitators rather than prescribers and learners would achieve language

proficiency by being repeatedly exposed to a number of native speaker's texts from which they would infer lexico-grammar patterns by themselves.

Flowerdew (2015) refers to three learning theories underpinning the DDL approach: "the noticing hypothesis, constructivist learning and Vygotskyan sociocultural theories" (p.16). The former is concerned with psycholinguistic processes behind learners' awareness of the differences between authentic language and the language produced by themselves. For instance, the noticing hypothesis could be used as the basis of a study in which learner and native corpora are provided to the participants for further examination. As for Vygotskyan theories, these state that knowledge is constructed between the learner and the teacher through collaborative dialogue, while not undervaluing the role of the teacher as a guide and facilitator. On the other hand, constructivist learning theories place learners as the centre of the picture and state that knowledge is only achieved by inductive processes in which the teacher has no relevance anymore.

During the last decades, mainstays of DDL have been confronted by several researchers, such as its inductive nature or the fact that direct and indirect use of corpora are understood as polarities. Concerning the latter case, Boulton (2012) stressed that there is a continuum between hands-on and hands-off use of corpora. Namely, that DDL can be practised by learners throughout access to corpora with different levels of guidance provided by teachers, in the form of planned DDL tasks, materials, or instruction for achieving a more autonomous learning while getting acquainted with corpora tools.

The notion of the continuum challenges both the abovementioned binary division between direct and indirect use of corpora as well as its allegedly inductive nature. Several of the reviewed research papers explored the matter, highlighting that through teacher-developed corpus-based materials or through teacher-directed tasks, learners would overcome possible setbacks when approaching corpora and their tools, such as the difficulty learners encounter to make inferences.

1.2. Typical features of DDL

In its most fundamental form, DDL tools would imply a corpus and a concordancer. For EFL teaching purposes, language is usually gathered in monolingual native corpora, being a corpus briefly defined as a "a large body of machine-readable texts" (Crystal, 2008, p.117), which can be 'raw' or annotated, i.e., provided with additional linguistic information.

A concordancer is a search engine that allows the learner to access and extract data from a given corpus. The data is presented in different forms, being the most common ones: a frequency list, which showcases words that appear most frequently within the corpus; concordance lines, featuring a search composed by a word or several words in their context, along with the rest of the sentence; collocations, words that statistically tend to appear together and n-grams, which showcase multiword clusters, usually from three to five words which appear frequently together in a corpus (Pérez-Paredes & Zapata-Ros, 2017).

1.3. Scope of the study

There is a lack of systematic reviews on DDL for language learning. In fact, Pérez-Paredes' (2019) systematic review of the 2011–2015 period in Computer-Assisted Language Learning (CALL) research is not only a rarity but also the newest contribution to the matter. His review explores research papers published by the five most relevant CALL journals: *CALL*, *CALICO Journal*, *LLT*, *ReCALL* and *System* during a period of five years.

DDL and language learning research papers represented only 4.2% of the total publications within these journals. Since all ICT-related research undergoes rapid changes in short periods of time, it appears relevant to explore whether the field has evolved during recent years.

The scope of this review is, thus, limited to research papers written during the last four years in five CALL-related journals, such as monographs, case studies, and reports on workshops that provide insights into data-driven learning used in EFL teaching and whose analysis ultimately aims to answer the following research questions in a structured and coherent way.

1.4. Research questions

1. What is the percentage of representation of DDL for EFL learning in the five most relevant CALL-related journals?
2. What are the trends amongst the selected DDL papers?
3. According to Pérez-Paredes (2019), there was a lack of theorisation on the role of corpora and DDL in second language learning in the research articles under analysis. Has this trend changed during the last four years?
4. Would a corpus-based analysis reveal undetected trends within the reviewed literature?

2. Methodology

In order to differentiate a systematic literature review from other traditional kinds also referred to as 'narrative reviews', Macaro (2020) stated that compared to the former the latter can be affected by "bias and lack of systematicity" (p.259). The means to reduce bias is to ensure there is transparency, "that a number of principles and procedures are adhered to from its very conception to the finished product" (p. 260).

The criteria to include a particular paper in this review are as follow: They needed to be original research papers, which means that book reviews or editorials were excluded from the analysis. They also needed to be published by the five CALL-related journals *CALL*, *CALICO Journal, LLT*, *ReCALL* and *System* during the years 2016-2019. They should be related to DDL applied to EFL, since it was decided to focus solely on English language learning, whereas in Pérez-Paredes' review, literature addressing DDL in all second language learning was included.

Therefore, the aim of this systematic review is not to exactly replicate Pérez-Paredes' work in a different span of time. In sum, his work was focused on DDL and language learning, the reviewed papers were published during a

period of five years (2010-2015) and his analysis was underpinned by Bax's framework of normalisation. Instead, Pérez-Paredes' review serves as a guideline by means of conducting a corpus-based analysis on the selected papers and by limiting the scope of the journals to be taken into consideration. Additionally, his findings were a starting point to explore noteworthy venues of the current literature .

2.1. Procedure

The systematic search was conducted by browsing particular keywords in the five CALL-related journals search engines, such as 'DDL', 'data-driven learning', 'EFL', 'ESP', 'corpora', and 'corpus', along with a filtered search for the period 2016-2019. Some hand searching was also done while navigating throughout the journals databases in order to ensure all significant research was selected.

Once the relevant titles were extracted, a preliminary reading of the papers excluded the ones which were concerned with DDL applied to the learning of languages other than English and research papers which were broader than expected, by approaching DDL only tangentially. The resulting selected papers were assigned an ID (see Appendix A) to later connect these IDs with particular categories.

According to Buckingham (2015), one of the first stages in every literature review involves the process of identifying "commonalities, connections and differences between the various ideas or concepts" (p. 98). In order to achieve this, a systematic mapping of the selected papers was conducted, retrieving information which was present in the vast majority of the studies. Macaro (2020) roughly defines the map as "a table which contains in very brief note form all the information needed to get an overall impression of the research."

Therefore, the following categories were included in a spreadsheet while attempting to answer RQ2 and RQ3: a) type of study, b) area of interest, c) theoretical framework, d) participants in the study, e) number of participants, f) further research, g) educational setting, h) participants' background i) participants' average English proficiency level, j) participants' L1, k) country, l) tools used by participants, m) time spent in instruction.

These categories varied while conducting a close reading of the existing research, except for ‘category c’ which is entirely dependent on RQ3 and therefore was included at the outset of this study. However, within the 13 categories, the last seven were not applicable in three of the selected papers, which did not involve participants’ direct use of corpora.

The outcomes of the systematic review will be related to the research questions and, while discussing these findings, expressing criticality. This is one of the flaws identified by some researchers who argue that novice writers usually fail to establish “a critical stance when writing literature review texts” (Bruce, 2014, p. 86). Following these guidelines, whenever the connection between the research questions in the selected papers and their findings is not clear or might seem biased, the issue will be addressed in the discussion section of the present literature review.

Concerning RQ4, a corpus-based analysis was conducted on the present sample which contained the 26 selected papers. As in Pérez-Paredes systematic review, Sketch Engine (Kilgariff, 2014) was used to extract multiword keywords that are representative of a particular corpus, i.e., at least two words that appear together in a corpus more frequently than in other, generally, a larger and broader corpus. Pérez-Paredes (2019) stated that this method was used “to uncover the scope and themes of the analysed papers” (p.11). To this end, the collection of papers was cleaned by removing irrelevant material like headers, figures, tables references, bibliographical lists, etc. (Leńko-Szymańska, 2017). In addition, the multiword keyword analysis was conducted on the sample corpus using the English Web 2015 corpus (enTenTen15) as a broader corpus reference.

3. Findings and discussion

3.1 Research question 1

The aim of RQ1 was to clarify what percentage of scholarly literature is related to DDL. As Table 1 shows, research on data-driven learning in the context of EFL teaching represented 3.36% of the published research during the years 2016-2019 in the five journals. Only 26 papers of the 772 published addressed the use of DDL for EFL purposes.

Table 1

Research published from 2016 to 2019 in the five CALL-related journals addressing DDL and corpora in EFL

Journal	Total	DDL/corpora in EFL	%
CALICO Journal	45	0	0
CALL	177	4	2.3
LLT	93	9	9.7
ReCALL	72	8	11.2
System	385	5	1.3
Total	**772**	**26**	**3.36**

ReCALL and *LLT* published the highest percentage of DDL-related papers. However, the latter published a special issue on corpora and language learning in 2017 which included eight of the total nine selected papers for further analysis.

As for the complete lack of DDL research published in *CALICO Journal*, it should be noted that the total of full original research papers were considerably lower than during the previous period tackled by Pérez-Paredes. In the latter case, an average of 22.6 papers were published per year, whereas during the 2016-2019 period the average was only 11.25 papers. After some research, it

appears that in 2015 *CALICO* moved to Equinox Publishing, described as "a young publishing house based in Sheffield, England" (Smith & Schulze, 2014, p. 4). However, It is uncertain whether this change is related to the decrease in its publication rate by almost a half. And yet to be explained why *CALICO Journal*'s DDL research falls from 5.3% during the previous period to 0% during the last four years.

In order to compare Pérez-Paredes' results with the current data, the five papers excluded from the present analysis for being focused on DDL and the learning of other second languages apart from English should be now considered. In this light, the percentage difference between both periods for all research related to DDL and second language learning is only 0.19%.

3.2 Research question 2

The purpose of RQ2 was to elucidate what trends arise from the analysis of the current literature on DDL and EFL learning. As it turned out, the vast majority of research papers —23 out of 26—were experimental or quasi-experimental studies, whereas the remaining three papers only involved a survey design and two of these were written by the same author with the same participants [ID 1 and 2]. In most of the current literature, however, researchers conducted a course which included a practical introduction to corpora, several tasks, and later analysed participants' use of corpora and, in some cases, their perceptions. Amongst these studies, the majority were preoccupied with measuring effectiveness on DDL, using pre-tests and post-tests in addition to other methods such as interviews with the participants, surveys, etc. Within this subdivision, experimental studies had at least two groups: usually, a control group with no access to corpora and an experimental group with access to corpora.

The fact that most of the research is effects-oriented may be explained by the lack of normalisation of DDL in mainstream teaching practice. Despite the enthusiasm of corpus linguists and corpora affordances, this approach is far from being integrated in the EFL classroom. Therefore, it is not surprising that most current literature is concerned with proving its effectiveness compared to other teacher-centred methods.

Regarding research's major areas of interest, the first one is academic writing, followed by L2 vocabulary learning. These results were expected, since DDL certainly allows learners' revision of their own writing production and several English for Academic Purposes (EAP) corpora are helpful tools which are available online. As for L2 vocabulary learning, it accommodates to corpora affordances, especially by the way of inferring meanings while examining words within their context. Interestingly, the latter studies heavily relied on guidance by teachers, who provided learners with glosses and tasks including edited paper-based concordances, thus, aligning with deductive approaches to DDL. Whereas studies concerned with academic writing encouraged autonomous learning and were more consistent with constructivist, deductive approaches to corpora use. Other related findings were that none of the research addressed corpus consultation to enhance listening skills, while just one paper tackled corpora and speaking skills, but only tangentially [ID 6].

Interestingly, some research was concerned with collaborative/individual access to corpora and deductive/inductive cognitive learning styles [ID 14, 24 and 25]. As stated in the introduction, the latter is a mainstay of DDL. Hence, it appears that recent literature attempts to challenge the role of teachers as mere facilitators within the data-driven learning approach. It is also worth noticing that unlearning a particular feature, a common mistake that takes into account learners' intralingual transfer, was the focus in one of the studies [ID 20]

With reference to participants' typology, the majority were students (22), followed by teachers (4) and pre-service teachers (2). The presence of teachers in DDL training programs is consistent with many authors' views, who have been long-time advocates for "the inclusion of corpora and CL in the syllabus of language teacher education programs" (Ebrahimi & Faghih, 2017, p. 121)"

With respect to the number of participants, it ranges from 327 to six, being the mean 108. As for the time spent on instruction, there was heterogeneity in measuring this value. Some researchers opted for measuring the time in hours, others in sessions or even whole semesters, with no specification as to how long the instruction on DDL lasted exactly. In rare cases, time spent in instruction was not specified. Finally, it was decided to divide this category in two: 'Workshop', for short courses that lasted less than five sessions and 'one-semester course', for the ones that lasted a whole semester, approximately 15 one-hour sessions.

The results show that both types of courses were evenly conducted. However, there was a particular study [ID 13] whereby the time spent in basic instruction lasted only three hours and was included in the classification as a workshop, even though the participants developed a task autonomously during twenty weeks. This task consisted of writing a term paper or a research proposal using online dictionaries an corpora tools without teacher feedback or further training. As it will be discussed later, the short duration of workshops was identified as a limitation by some researchers who called for studies in which the training lasted longer to strengthen the results obtained.

Regarding the educational setting where studies were conducted, only three of them were set in the context of non-higher education. However, it is promising that some attempts are being made to introduce DDL in secondary education, where the difficulties are more severe than in higher education. Teachers are not usually acquainted with corpora tools, syllabuses are prepared in advance and do not admit much innovation, particularly in state-funded schools. On the contrary, researchers are also teaching at universities, which gives them the opportunity to implement DDL in their classes.

As for participants' background, only the ones who were students in higher education were taken into account. Interestingly, this value ranges from sciences to humanities and, contrary to what it might be expected, only four of 19 studies in which this value was applicable were exclusively students of English and Humanities. Research papers including students of science-related disciplines accounted for a total of seven. In four of these cases, participants were graduate students who were preparing themselves to write a term paper or a research proposal in English, thus concerned with academic writing. Nevertheless, there is not a correlation between participants studying science-related disciplines and participants' qualification, since in three of the research papers, participants were undergraduates majoring in science-related disciplines.

It was found pertinent to ask whether the mentioned three studies have any commonalities which might explain why these science, technology, engineering and mathematics (STEM) students are using data-driven learning in the context of EFL at undergraduate level. The main feature that these three studies share is that the DDL course was conducted at East Asian universities (Japan and Korea). Even though it is not specified, it is likely that these students

were enrolled in university programs where English was used as medium of instruction, hence the necessity of students to improve on their writing performance. However, this matter will be addressed again later.

Concerning participants' average English proficiency level, some of the studies used the Test of English for International Communication (TOEIC) and the International English Language Testing System (IELTS) amongst other standards. Therefore, it was decided to convert these different scores into the Common European Framework of Reference for Languages (CFER) standards. A complete range of levels, with the exception of C1, is featured in the selected papers in which this value is applicable, being B2 the most repeated proficiency level in the sample. With reference to the presence of C2 level, a remark is worth being outlined: C2 level is not really representative, even though it had to be included since seven English native speakers participated in one of the studies along with non-native speakers [ID 7].

Another mainstay of data-driven learning is challenged since, contrary to the belief that this approach is more suitable for advanced English learners, there is an attempt to teach DDL to low-level learners of English with promising results [IDs 8,17, 19, 20, 21 and 25], Furthermore, two of these studies [ID 17 and 20] not only contributed to pioneer the field of low-level learners response to DDL, but also conducted their analysis in a non-higher educational setting, which is indeed a novelty amongst DDL research.

With respect to participants' L1 and countries where the studies were conducted, it is found that the majority were set in East Asian countries (15) – including Vietnam as part of this region– with Korean and Chinese as participants' mostly spoken mother tongues, followed by US and Canada (4) with a variety of L1s spoken by the participants. Lastly, only three studies were conducted in Europe. Different reasons might explain this trend wherein most of the studies were run in this particular area. In this note, Karras (2016) is confident about DDL implementation within the Asian context and states that it could be achieved "relatively painlessly by a fairly small number of dedicated L2 educators who have received relevant DDL vocabulary training and instruction" (p. 182). Further analysis would be necessary, although another possible explanation might lie on students having better access to computer labs in East Asian schools and universities. Besides, culture-related views on technology and education, like

mobile learning as a helpful tool in the classroom, may encourage their use. On the contrary, mobile devices are far from being accepted in educational contexts in Europe and they are rather perceived as a distraction in the classroom.

Digital skills are also important to consider when introducing DDL to students. In this regard, South Korea is listed by UNESCO as one of the three leading countries in highest levels of digital skills of children (Fau, 2018). This could be one of the reasons behind the fact that Korean participants were involved In as many as six of 19 studies in which this value was applicable or specified. A different rationale behind these results might be the increasing number of universities in East Asia which adopted English as a medium of instruction. An example would be Hong-Kong, with six of its eight government-funded universities being English medium (Kirkpatrick, 2014).

Entire continents such as Africa and great areas like Latin America are not represented at all in the selected papers. This matter could also be connected to the lack of digital skills within their population, a consequence of economic factors reflected in the lack of available resources for learners. This is especially the case of Africa where, according to recent stats (ITU, 2019), the rate of access to the Internet is the lowest in the world. Accessing to the Internet is relevant not only because it enables learners to continue their training and tasks at home in a more autonomous style, but also to develop their digital skills.

Finally, as for the data concerning which corpora tools were used by participants, they were divided into the two following groups and their frequency of use was registered:

1. Corpora
    a. General corpora
    b. Specialized corpora
        i. Discipline-oriented
        ii. Graded corpora
    c. Ad-hoc corpora
2. Concordancers
    a. Available concordancers
    b. Ad-hoc concordancers

The most frequently used general corpus was the Corpus of Contemporary American English (COCA). A general corpus is defined as a large collection of

texts. COCA, in particular, comprises more than 1bn words. These collections of texts may belong to different registers, text types and subject fields. COCA was followed in use by the British National Corpus (BNC). Other present-day American English corpora such as the Brown Corpus and the Open American National Corpus (OANC), together with COCA, slightly outnumbers British English corpora use. Nonetheless, American English general corpora were not used alone except for two studies conducted in Canada and Taiwan [IDs 13 and 24].

Within the specialized corpora which, as opposed to general corpora, is genre-specific, the most frequently used discipline-oriented corpus was Google Scholar, followed by both the Michigan Corpus of Academic Spoken English (MICASE), and the Michigan Corpus of Upper-Level Student Papers (MICUSP).

A parallel corpus, a Japanese-English corpus called WebParaNews, was used in two studies. Boulton (2017) stated that a parallel corpus is still rare outside specialized translation courses — none of which are featured in this review — or outside courses designed for learners at lower levels of second language proficiency. In the latter case, this trend continues, since the only two studies whose participants used a parallel corpus [IDs 19 and 25] were at the lowest level of proficiency, being in both cases A1-A2 learners of English.

Several researchers built their own corpora. In two studies, a learner corpus was compiled for raising awareness on common mistakes made by non-native speakers. One paper's main concern was improving students' writing [ID 5] and the other addressed how to unlearn a particular incorrect structure by comparing a native corpus with a learner corpus [ID 20].

The Hong Kong Graduate Corpus, an ad-hoc annotated and genre-specific corpus, was built by researchers for students' use [ID 15] and included theses from several disciplines. Another annotated corpus with genre conventions was built to focus students' attention into the macro and micro-structures present in Academic English texts [ID 7]. Both studies used DDL in order to improve students' academic writing.

In annotated corpora, different linguistic information may be added manually or automatically via Natural Language Processing (NLP). Corpus annotation is commonly perceived amongst the CL community as an "activity that enriches and gives 'added value' to a corpus" (Hovy & Lavid, 2010, p.26).

Considering that researchers also had to perform their own analysis, corpus annotation for students' use is certainly an ambitious task that has been completed by only a few researchers.

Regarding concordancers, some students were trained in using ready available ones which could be used either online or by downloading a software application. Most concordances were also built ad-hoc for teaching purposes, followed by AntConc and LexTutor. In each study where learners were provided with an ad-hoc corpus, researchers also developed their own concordance tool. However, in some rare cases, a specific concordancer which obtained its data from a variety of online corpora was also developed.

Considering all 26 papers, further research was encouraged on the development of longitudinal studies whereby post-delayed tests could be taken by the participants to explore the differences in knowledge retention and/or to secure DDL approaches effectiveness over time. In the same line, researchers also called for larger-scale studies involving a larger number of participants. Three of these studies highlighted the need for more time to be spent on instruction as well as truly experimental conditions to perform their research. And, finally, in two cases, authors' proposal for further research was to develop a model in teacher training programs which included corpora use in the EFL classroom. Appendix B contains several figures illustrating the trends that emerged from the analysis of these papers.

3.3. Research question 3

In Pérez-Paredes' systematic review (2019), a lack of theoretical support in CALL research was noticed and a need for further theorisation was established to favour a better understanding on how second language learning theories are connected with empirical evidence from DDL approaches to teaching. Flowerdew (2015) had also previously noted that learning theories were scarcely discussed in depth in scholarly literature. Consequently, it was relevant to the present review to elucidate whether recent research accommodate to this view or whether this trend has changed during the last four years. After the collection of papers was examined, it was found that lack of theorisation persists in DDL-related research.

The general tendency was that, since the majority of papers were of experimental nature, they heavily relied on similar research in their literature review sections, without tackling second language learning theories and their relationship with data-driven learning. For this reason, researchers had several foci which were mainly related to procedurals in methodology or the designing of the lessons (e.g., Flowerdew's steps in ID 20 and 23)

Yet, nine studies referred to a number of theories, moreover, Vygotskyan constructivist theories [ID 3, 15 and 16]. Several authors also expressed their willing to bridge the gap between practice and theory, even though this matter is not later addressed nor developed further in their findings. (See Appendix C for more details).

In contrast with the overall orientation within the current literature, a particular paper [ID 7] not only conducted an experimental effect study whose participants' used an ad-hoc annotated corpus, a designed for the occasion concordancer and a custom-built Automated Writing Evaluation (AWE) tool, but its findings were discussed while taking into consideration language learning theories. Hence, it could be argued that researchers' successfully attempted to theoretically underpin their own empirical analysis with a "DDL-substantiated conceptual framework" (Cotos, Link, & Huffman, 2017, p. 124).

3.4. Research question 4

Lastly, in RQ4 it was questioned whether a corpus-based analysis would reveal undetected trends within the current literature. Sketch Engine was used to conduct a multi-word analysis on a small-scale purpose-built corpus to further examine the body of research. Firstly, a hand-curated version of the research papers was used as the focus corpus, containing a total of 174.505 words. It was later uploaded in individual text files in order to verify whether multi-words are distributed evenly across the focus corpus, i.e., that a particular combination of two words appeared together in a number of papers. The English Web 2015 corpus (enTenTen15), containing over 15bn words, was used as the reference corpus.

The multi-word analysis results support the findings of this study. Amongst the top 15 words appearing together in the focus corpus, the most relevant are:

'experimental group', 'vocabulary learning', 'learner corpus', 'academic writing', 'vocabulary acquisition' and 'error correction'. However, it is not found that it particularly serves to the purpose of uncovering trends which were not noticed while conducting the non-corpus research. The complete list with the top 50 multi-words can be found in Appendix D for closer examination.

It should be noted that too broad pairs such as 'data-driven learning' or 'corpus use' and multi-words distributed unevenly, such as 'glossary information', which only appears in two papers, have been ignored. To illustrate the latter case, an uneven and an even distribution can be compared in Figure 1 and Figure 2, extracted from Sketch Engine.

Figure 1

Distribution of hits in the focus corpus for the pair 'academic writing'

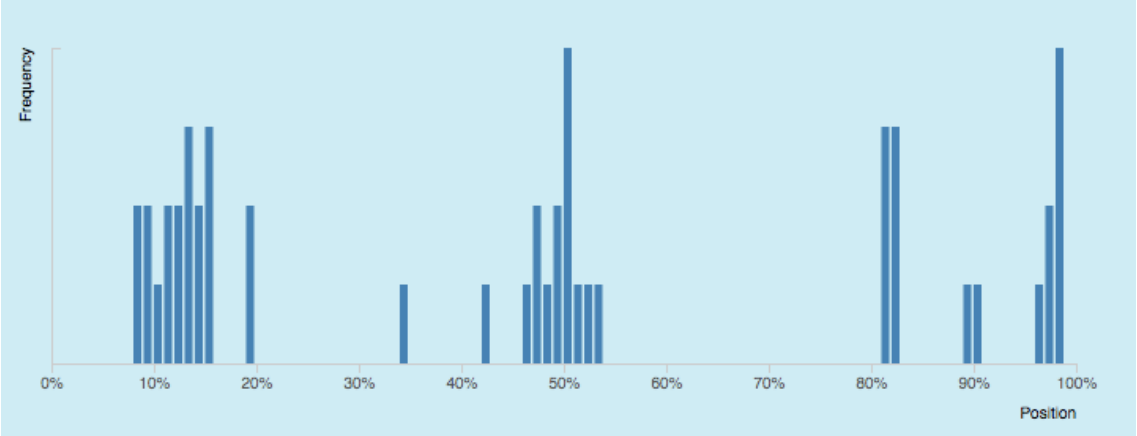
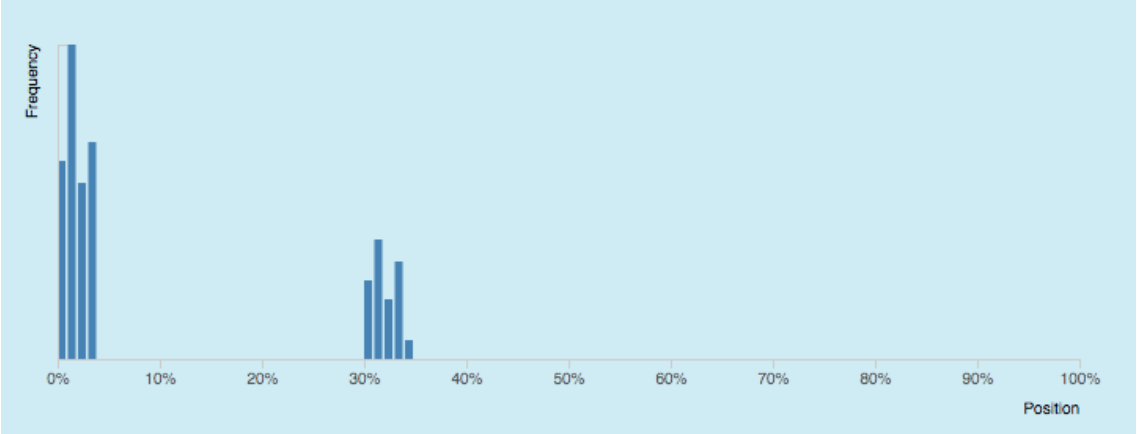


Figure 2

Distribution of hits in the focus corpus for the pair 'glossary information'



4. Conclusion

A total of 26 papers were reviewed in this study. The findings of this systematic review highlighted that little research on DDL for EFL learning was published in the five CALL-related journals during the last four years, accounting for only 3.4% of total publications.

The great majority (23) of the selected papers were of experimental or quasi-experimental nature whereby results were analysed after a course on DDL was conducted by the same researchers. Moreover, effect studies which are either concerned with the effectiveness of DDL compared to other more traditional teacher-centred approaches or with the effectiveness of different of learning styles in a DDL classroom. As examples of the latter, three papers attempted to clarify which approaches were more suitable for learners' use of corpora between deductive and inductive learning and between individual and collaborative learning

The number of participants ranged from 327 to six. As for participants' typology, they were mostly undergraduates studying in different fields, from humanities to science. The bulk of the research focused on enhancing participants' writing skills, particularly their academic writing. In addition, the tendency was to encourage lexico-grammatical awareness amongst the learners.

In terms of time spent in instruction, approximately half of the experimental research was classified as a workshop and the other half as a one-semester course. The majority of the experiments were conducted in East Asian countries (11), most notably in South Korea.

Regarding trends in the use of corpora tools, COCA was the most used general corpus. As for specialized corpora, the majority of the participants used Google Scholar. In many cases, researchers built an ad-hoc corpus, as well as their own concordancers, being AntConc the most used concordancer amongst the ready available ones.

Researchers were not able to develop, neither analyse post-delayed tests. As a result, the commonest call for further research identified in the collection of papers was the necessity of conducting more longitudinal studies. This particular issue is in line with the evidence found both in Pérez-Paredes work and in the present review, wherein current literature on DDL lacks theorisation and is mostly

concerned with research methodology, optimising experiment design and procedurals.

It appeared that conducting a multi-word analysis in Sketch Engine was not helpful to reveal undetected trends in DDL research, even though the results supported the evidence found by the non-corpus based approach. One of the perceived limitations of this tool was that similar spellings, such as British English and American English spellings were overlooked, hence, some pairs did not appear in the frequency list. Additionally, since multi-word analysis was not able to establish relationships amongst words with similar meanings, further literature reviews could conduct a corpus-based analysis on a larger body of papers using more sophisticated techniques within the NLP field, such as latent semantic analysis.

The main contribution of this study was to demonstrate that there is still little research on DDL applied to EFL learning. In addition, the uses of corpora in current teaching scenarios were explored in depth and trends were successfully identified. Lastly, after conducting a multi-word analysis, a different strategy was proposed to achieve more effectiveness in further systematic reviews.

5. References

Aijmer, K. (2009). *Corpora and language teaching.* Amsterdam: Benjamins.

Boulton, A. (2012). What data for data-driven learning? *EuroCALL Review*, *20*(1), 23–27.

Boulton, A. (2017). Data-Driven Learning and Language Pedagogy. In *Language, Education and Technology: Encyclopedia of Language and Education.* https://doi.org/10.1007/978-3-319-02237-6

Bruce, I. (2014). Expressing criticality in the literature review in research article introductions in applied linguistics and psychology. *English for Specific Purposes*, *36*(1), 85–96. https://doi.org/10.1016/j.esp.2014.06.004

Buckingham, L. (2016). *Doing a research project in English studies a guide for students*.

Cotos, E., Link, S., & Huffman, S. (2017). Effects of DDL technology on genre learning. *Language Learning and Technology*, *21*(3), 104–130.

Crystal, David. (2008). *A dictionary of linguistics and phonetics*. Malden, MA ; Oxford : Blackwell Pub

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

Fau, S., & Y. M. (2018). *Managing tomorrow's digital skills: What conclusions can we draw from international comparative indicators?* (Rep. No. 0000261853). UNESCO.

Flowerdew, L. (2015). Data-driven learning and language learning theories. In A. Boulton & A. Leńko-Szymańska (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 15–36). Amsterdam: John Benjamins Publishing Company.

Gilquin, G., & Granger, S. (2012). How can DDL be used in language teaching? In A. O'Keeffe & M. McCarthy (Authors), *The Routledge handbook of corpus linguistics* (pp. 359-369). Abingdon: Routledge.

Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, *22*(1), 13–36.

Huang, Z. (2014). The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL*, *26*(2), 163–183. https://doi.org/10.1017/S0958344014000020

ITU. (2019). *Measuring Digital Development: Facts and figures* (Rep.). Geneva, Switzerland: International Telecommunication Union. doi:https://www.itu.int/en/ITUD/Statistics/Documents/facts/FactsFigures2019.pdf

Johns, T. (1991). Should you be persuaded: two examples of data-driven learning. *ELR Journal 4: Classroom Concordancing. Birmingham: CELS, The University of Birmingham*, 1–16.

Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL*, *28*(2), 166–186. https://doi.org/10.1017/S0958344015000154

Kirkpatrick, A. (2014). English as a Medium of Instruction in East and Southeast Asian Universities. *The Asian Journal of Applied Linguistics*, 15–29. https://doi.org/10.1007/978-94-007-7972-3_2

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

Macaro, E. (2020). Systematic reviews in applied linguistics. In J. McKinley & H. Rose (Authors), *The Routledge handbook of research methods in applied linguistics* (pp. 254-269). London: Routledge, Taylor and Francis Group.

McEnery, T., & Xiao, R. (2011). *What corpora can offer in language teaching and learning. In E. 191 Hinkel (Ed.), Handbook of research in second language teaching and learning*. (2007), 364–380.

Pérez-Paredes, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, 1–26. https://doi.org/10.1080/09588221.2019.1667832

Pérez-Paredes, P., & Zapata-Ros, M. (2017). Patrones de pensamiento computacional y corpus lingüísticos: el aprendizaje de lenguas con datos lingüísticos. Retrieved from https://core.ac.uk/download/pdf/154903776.pdf

Smith, B., & Schulze, M. (2014). Publishing call research in volatile times. *CALICO Journal*, *31*(3), i–v. https://doi.org/10.11139/cj.31.3.i-v

Talai, T., & Fotovatnia, Z. (2012). Data-driven learning: A student-centered technique for language learning. *Theory and Practice in Language Studies*, *2*(7), 1526–1531. https://doi.org/10.4304/tpls.2.7.1526-153

Appendix A

List of the selected research papers

Authors	ID	Title	Year	Journal
Ballance	1	Analysing concordancing: a simple or multifaceted construct?	2016	CALL
Ballance	2	Pedagogical models of concordance use: correlations between concordance user preferences	2017	CALL
Crosthwaite	3	Retesting the limits of data-driven learning: feedback and error correction	2017	CALL
Pérez-Paredes et al..	4	Language teachers' perceptions on the use of OER language processing technologies in MALL	2017	CALL
Ackerley	5	Effects of corpus-based instruction on phraseology in learner English	2017	LLT
Bardovi-Harlig, et al.	6	The effect of corpus-based instruction on pragmatic routines	2017	LLT
Cotos et al.	7	Effects of DDL technology on genre learning	2017	LLT
Hadley & Charles	8	Enhancing extensive reading with data-driven learning	2017	LLT
Shin & Han.	9	Teaching Google search techniques in an L2 academic writing context	2017	LLT
Hansol et al.	10	The effects of concordance-based electronic glosses on L2 vocabulary learning	2017	LLT
Leńko-Szymańska	11	Training teachers in data-driven learning: Tackling the challenge	2017	LLT

Li	12	Using corpora to develop learners' collocational competence	2016	LLT
Yoon	13	Concordancers and dictionaries as problem-solving tools for ESL academic writing	2016	ReCALL
Cho	14	Task dependency effects of collaboration in learners' corpus consultation: An exploratory case study	2019	ReCALL
Crosthwaite et al.	15	Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning	2017	ReCALL
Ebrahimi & Faghih	16	Integrating corpus linguistics into online language teacher education programs	2016	ReCALL
Karras	17	The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam	2019	ReCALL
Hansol et al.	18	Advancing CALL research via data-mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment	2016	ReCALL
Mizumoto et al.	19	Development of a scale to measure learners' perceived preferences and benefits of data-driven learning	2018	ReCALL
Moon & Oh	20	Unlearning overgenerated be through data-driven learning in the secondary EFL classroom	2015	ReCALL

Mueller & Jacobsen	21	A comparison of the effectiveness of EFL students' use of dictionaries and an online corpus for the enhancement of revision skills	2019	System
Chen et al.	22	Introducing in-service English language teachers to data-driven learning for academic writing	2017	System
Larsen-Walker	23	Can Data Driven Learning address L2 writers' habitual errors with English linking adverbials?	2019	System
Lee & Lin	24	The effect of the inductive and deductive data-driven learning (DDL) on vocabulary acquisition and retention	2016	System
Mizumoto & Chujo	25	Who is data-driven learning for? Challenging the monolithic view of its relationship with learning styles	2019	System
Pérez-Paredes et al..	26	Mobile data-driven language learning: Affordances and learners’ perception	2019	System

Appendix B

Figures on general trends within DDL research

Figure B1

Major areas of interest

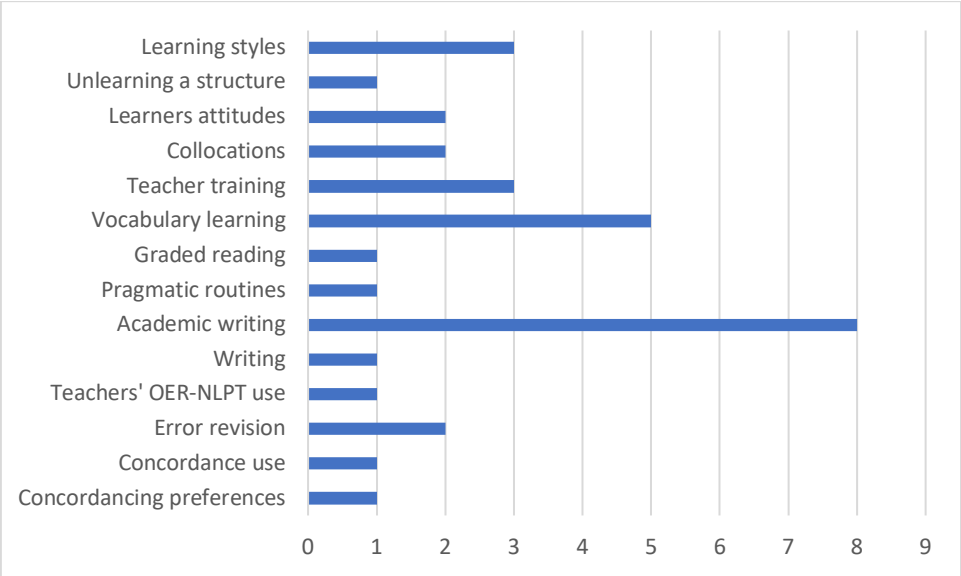


Figure B2

Time spent in instruction

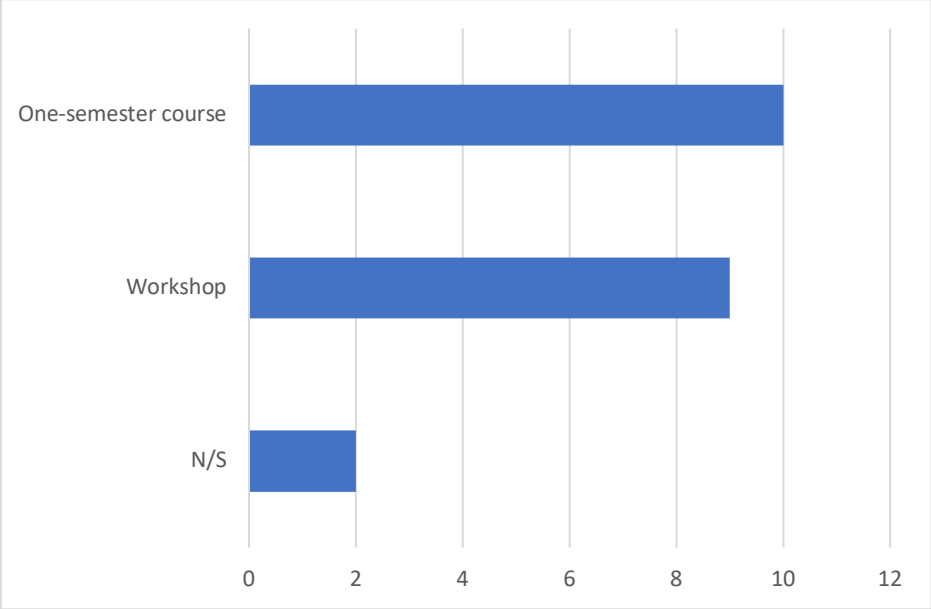


Figure B3

Educational setting of experimental studies

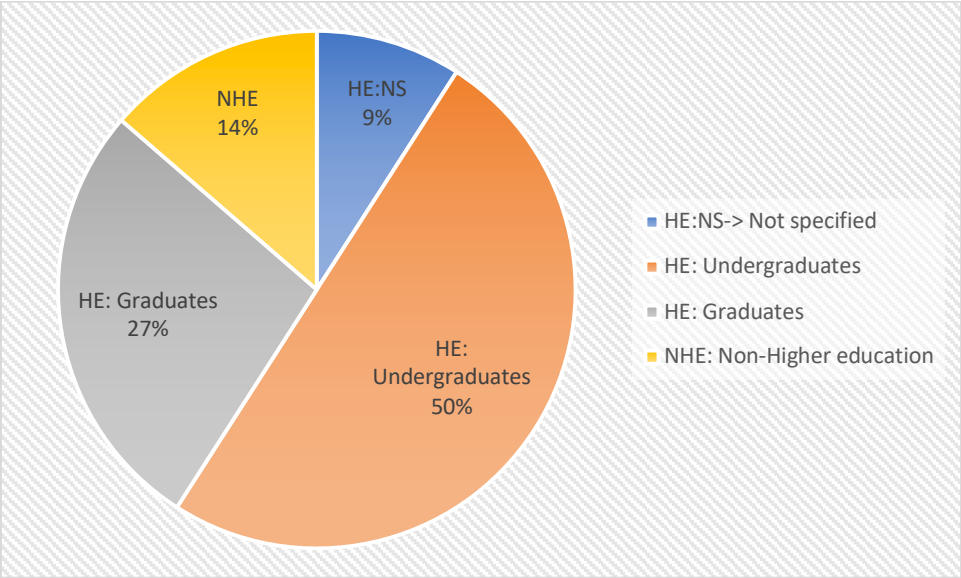


Figure B4

Countries where experimental research was conducted

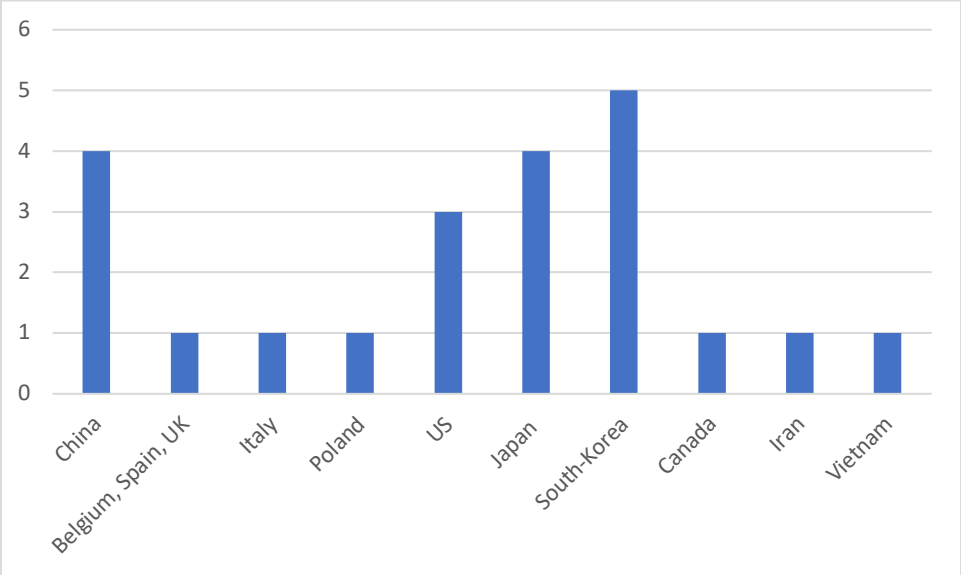


Figure B5

Use of general corpora

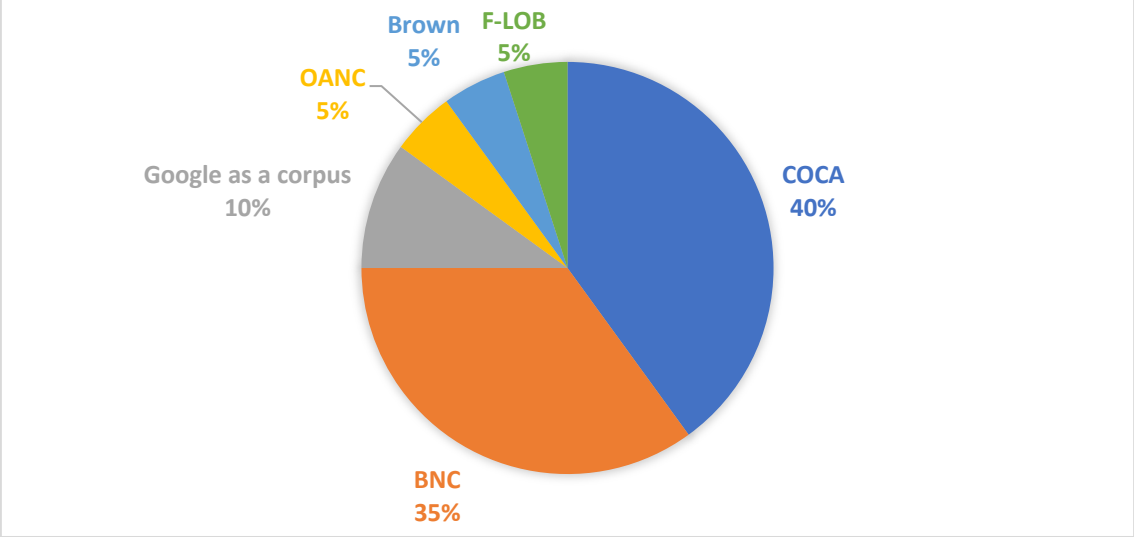
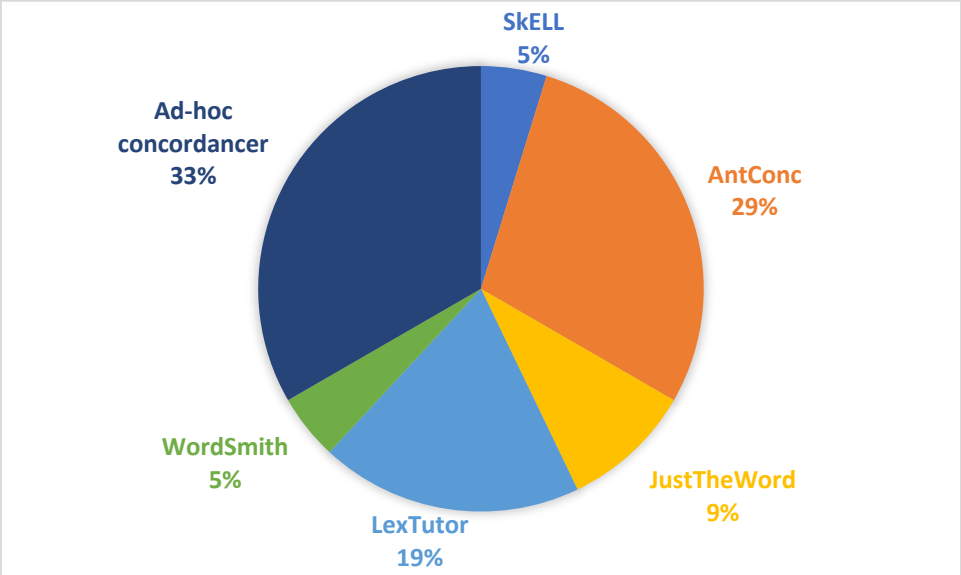


Figure B6

Use of concordancers



Appendix C

Research that referred to learning theories

ID	Type of study	Area of interest	Theories mentioned
2	Survey	Concordance use	Theoretical perspectives on learner concordance use
3	Exploratory study	Error revision	Vygotskian constructivist theories
7	Experimental effect study	Academic writing	Cognitive writing theories: the knowledge-telling/knowledge-transformation model of Bereiter and Scardamalia (1987)
8	Experimental effect study	Graded reading	Personal construct theory
10	Experimental effect study	Vocabulary learning	The noticing hypothesis
13	Effect and exploratory study	Academic writing	Cognitive extension vs cognitive distraction
14	Exploratory experimental effect study	Collaborative learning	1. Conceptual and procedural competence 2. Intersubjectivity 3. Power inequality
15	Exploratory study	Academic writing, learners' patterns of corpus usage	Vygotskian constructivist theories
16	Exploratory study	Teacher training, virtual learning environment	Vygotskian constructivist theories
25	Exploratory experimental effect study	Learning styles: inductive/deductive learning	Aptitude-Treatment Interaction

Appendix D

Top 50 multi-words found in the focus corpus

Term	Score	Freq	Ref freq	Rel freq	Rel ref freq
corpus consultation	781.450	167	14	780.450	0.001
glossary information	304.460	65	19	303.768	0.001
data-driven learning	280.280	60	84	280.401	0.005
concordance use	276.730	59	0	275.728	0.000
corpus use	276.730	59	11	275.728	0.001
experimental group	250.480	70	5708	327.135	0.310
vocabulary learning	192.090	42	513	196.281	0.028
learner corpus	172.360	37	167	172.914	0.009
academic writing	168.990	49	6649	228.994	0.362
vocabulary acquisition	161.180	36	937	168.241	0.051
control corpus	155.220	33	0	154.221	0.000
expert corpus	155.220	33	0	154.221	0.000
procedural task	155.060	33	31	154.221	0.002
error correction	142.120	43	7759	200.954	0.422
language teaching	137.970	49	12279	228.994	0.668
direct corpus	117.830	25	0	116.834	0.000
learner use	113.050	24	36	112.161	0.002
language learning	112.830	48	18340	224.321	0.997
immediate post-test	108.380	23	31	107.487	0.002
corpus query	108.270	23	46	107.487	0.003
vocabulary knowledge	104.820	23	655	107.487	0.036
collocational use	103.810	22	0	102.814	0.000
different learner	103.190	22	122	102.814	0.007
authentic language	101.780	22	377	102.814	0.021
language use	99.150	29	6951	135.527	0.378
online corpus	98.940	21	48	98.140	0.003
experimental class	98.160	21	185	98.140	0.010
word form	94.420	21	931	98.140	0.051
corpus analysis	92.430	20	421	93.467	0.023
b2 learner	89.790	19	0	88.794	0.000
word use	87.520	19	482	88.794	0.026
language awareness	86.590	19	691	88.794	0.038
present study	86.300	75	56519	350.502	3.074
collocational competence	85.120	18	7	84.120	0.000
conceptual task	85.030	18	27	84.120	0.001
target word	82.480	18	598	84.120	0.033

reference resource	81.140	18	906	84.120	0.049
target vocabulary	79.730	17	169	79.447	0.009
inductive approach	78.260	17	520	79.447	0.028
target language	77.240	24	8554	112.161	0.465
current study	76.130	33	19111	154.221	1.039
partial eta	75.700	16	34	74.774	0.002
query syntax	74.000	16	457	74.774	0.025
control group	73.000	60	52518	280.401	2.856
inductive group	71.100	15	0	70.100	0.000
inductive learning	70.400	15	189	70.100	0.010
speech act	69.900	16	1547	74.774	0.084
effect size	68.650	19	5673	88.794	0.308
second language	68.490	45	38348	210.301	2.085