Reserved for Research? Normalising Corpus Use for School Teachers

Rachel Allan (Mid-Sweden University)

Abstract
There has been much discussion about the persistent gap between research and practice in the use of corpora in the classroom (Frankenberg-Garcia 2012; Chambers 2019 among others), despite strong evidence of its benefits (Boulton 2017). The majority of studies into data-driven learning (DDL) have been carried out by those with a particular interest and skill level, predominantly in higher education, and the need to complement these with a broader base of studies involving practising language teachers in a school environment has been highlighted (e.g., Boulton 2010; Chambers 2019). For such studies to take place, however, more school teachers need to be made aware of DDL and its potential for use in the classroom.

This article discusses what we can learn from research into DDL with younger learners and teacher training in this context in order to shape a teacher training programme. It describes a pilot project introducing DDL to a group of secondary school student teachers (STs) of English at a Swedish university, and their responses to it regarding the feasibility of including it in their future teaching practice. The need for further training, particularly in practical pedagogical applications suitable for their learners, was apparent, echoing the outcomes of previous studies. It is suggested that integrating a range of classroom-focused DDL activities throughout their remaining course may be an effective approach. This also provides an opportunity to raise awareness of pre-prepared resources and novel approaches to DDL more likely to appeal to their learners, and practical examples of this are discussed.

Keywords: data-driven learning; teacher training; secondary school; classroom applications

1. Introduction

The gap between research and practice in the use of corpora in the classroom is widely acknowledged (Frankenberg-Garcia 2012; Chambers 2019 among many others). While a broad range of studies in the field indicate that there have been advances in classroom corpus use at tertiary

Allan, Rachel. 2023. 'Reserved for Research? Normalising Corpus Use for School Teachers.' *Nordic Journal of English Studies* 22(1): 68–92.

level, and a strong bank of evidence of the effectiveness of data-driven learning (DDL) has been amassed (Boulton 2017; Boulton and Vyatkina 2021), there has been much less attention paid to corpus use in the primary and secondary school classroom. As Chambers (2019: 464) points out, the majority of studies into DDL have been carried out by those with a particular interest and skill level, predominantly in higher education. Although there is evidence of increasing interest in incorporating DDL into the mainstream classroom (e.g., Crosthwaite 2019), effective approaches to teacher training will be required to achieve this. There have been developments in corpus training in teacher education courses (e.g., Farr 2008; Leńko-Szymańska 2014; Naismith 2017; Zareva 2017; Callies 2019 among others), but the majority of these have concerned teachers of adult students, rather than those in a school context. Also, while such studies suggest that teachers are positive towards corpus use, there is little evidence to date that this leads to corpus use in their teaching practice.

This paper describes a pilot course introducing DDL in an online teacher training programme for secondary school teachers of English in Sweden, aiming to encourage its normalisation in the classroom (see Chambers 2019: 461 for a discussion of normalisation). Empirical studies of DDL with younger learners and DDL in teacher training courses are first considered, together with their implications for the pilot. The pilot is then described and evaluated. Ways in which DDL could be embedded into further modules of this course are proposed, with an emphasis on classroom applications for younger learners, thus giving the trainees practical tools to use.

2. Previous research

Data-driven learning, used here with the broad definition of 'using the tools and techniques of corpus linguistics for pedagogical purposes' (Gilquin and Granger 2010: 359), can no longer be seen as a new approach. It has been in existence for more than 30 years, and has attracted a large number of studies into its use in that time (see Boulton and Vyatkina 2021 for a recent comprehensive review). As noted above, DDL appears to have gained some ground at tertiary level, but there is still little evidence of corpus awareness or use in mainstream schools (see, for example, Callies' (2019) follow up survey to Mukherjee (2004)). However, studies examining pedagogical applications of corpora with younger learners are increasing, as noted by Boulton and Vyatkina (2021) in their review.

Although they report that only 9% of the total 477 DDL studies they reviewed indicated an institutional context with younger learners, studies from the 2016 to 2019 period included a higher proportion with a focus on younger learners. Half of the 32 studies involving high school and middle school students were published in this period, some of which are discussed below.

2.1 DDL and younger learners

There has been some ambivalence regarding the use of DDL with younger learners, with studies highlighting the importance of teacher mediation and appropriate training in this context (e.g., Kaltenböck and Mehlmauer-Larcher 2005; Braun 2007). Braun (2007: 325) points out the complexity of integrating DDL into the school context, noting that this 'concerns both the corpus content (ensuring appropriateness and complementarity) and the ways in which the corpus is explored and exploited'. She advocates for a combination of classic corpus-linguistic methods and other methods, and points out that this creates a methodological challenge that teachers need to be trained for. More recent studies echo this need for mediation of the corpus. Frankenberg-Garcia's (2014) study found that Portuguese high-school students of English benefited from corpus input, but attributed this to the fact that 'the learners were essentially spoon-fed with the right type of examples' which they were unlikely to find through direct interface with the data. However, this may depend as much on proficiency level as age. Soruç and Tekin (2017) report positive outcomes, both in terms of learning effects and attitudes, in their study with Ugandan secondary school students using DDL autonomously with the British National Corpus (BNC), but the researchers consider that the high proficiency level of their students may have contributed to this.

Different approaches to mitigating the complexity of the corpus and the corpus interface have been considered. Mirzaei et al. (2015) explored the use of teacher-designed software, LexisBOARD, in an attempt to provide a more appropriate interface. Although results were encouraging, how far this may be transferable to other contexts is questionable given the many variables to be considered when creating corpus-based activities (see O'Keeffe et al. 2007). Another approach to moderating the interface between learners and corpora has concerned the type of corpora used. Graded corpora (i.e., corpora comprising limited and manageable text sources such as graded readers) have been found to offer a useful

alternative for learners at lower levels, limiting exposure to complexity while retaining sufficient authenticity of language (Allan 2009). Learner corpora have also been used, for example in Moon and Oh's (2018) study with Korean high school students, who used a learner corpus and a self-compiled graded corpus to explore the overuse of *be* with Korean high school students. They found that although DDL had a positive effect on learning and motivation for many of the students, some learners struggled with interpreting the concordance data even when graded.

The inductive approach typically used in DDL has been found to be an issue in some contexts. Moon and Oh (2018) found that some students were demotivated by this, and the researchers emphasized the importance of the scaffolding role of the teacher. Instruction style was also highlighted as an issue in Szudarski's (2019) study with lower-level Polish secondary school students, who were taught non-transparent formulaic language using paper-based DDL. This led to modest gains in knowledge, and although there was a broadly positive attitude to the approach, these learners were more positive towards more traditional deductive teaching approaches. This, and their limited proficiency, proved to be the main challenges to overcome, leading Szudarski to echo others (e.g., Wicher 2019; Boulton and Vyatkina 2021) in acknowledging the need to consider the local context, and apply DDL in an appropriate way.

Other studies have emphasized time demands of DDL. In a study into using DDL for vocabulary acquisition in an international high school, Karras (2016) concluded that DDL is a viable approach for lower L2 level secondary students as long as sufficient student training time is allowed. Beyond the initial learning curve, the approach was not found to be difficult by the majority of these learners (Karras 2016: 182). Although this seems promising, further research needs to be done to quantify the time this may take, and the viability of teachers investing this time both in their own training and the training of their learners.

A recent positive trend has been the use of more innovative approaches to DDL to make it more appropriate to younger learners. The focus of this tends to be on using multimedia corpora to add additional support and increase motivation. Applications include using pedagogical multimedia corpora, e.g., using the SACODEYL or ELISA corpus (Pérez-Paredes 2019) where there is an audio-visual element as well as the text. Creative use of more general, non-pedagogical applications is also a possibility, as Meunier (2019) points out, with her suggestions for using a

range of tools, such as the PlayPhrase.me app or web interface, which generates videoclips from films or series in response to a search for a word or phrase. YouGlish similarly offers the same function for YouTube videos (Meunier 2020). Another suggestion is to 'DDL-ize' LyricsTraining (Meunier 2019), an application which can be used to create gap-fill exercises for language learners while watching and listening to music videos. However, little empirical research for such approaches is available at this stage.

The main message that emerges from these studies is the need for mediation by teachers to use DDL effectively with younger learners. This in turn requires teachers to be sufficiently knowledgeable about DDL to select appropriate interfaces and corpora, and to be able to construct effective tasks while meeting the requirements of the syllabus (Leńko-Szymańska 2014: 272). They also need to meet the specific needs of their context, regarding issues like instruction style, time constraints, and access to technology, particularly high bandwidth multimedia applications. Considering the demands this puts on them, teachers need to be convinced of its value, placing considerable importance on how they are introduced to and trained in DDL.

2.2 Teacher education in DDL

Findings from corpus training in teacher education courses have been generally positive, with DDL found to be positively perceived by trainees, and useful for increasing language awareness (Farr 2008; Leńko-Symańska 2014; Zareva 2017; Callies 2019). However, there is a growing awareness of the time investment required for teachers to become competent users of DDL; increasingly, a distinction is being drawn between developing technical and corpus manipulation skills for language awareness purposes, and the skills required to develop appropriate pedagogical applications (Leńko-Szymańska 2017). The majority of studies have related to pre-service courses for language teachers of adults, i.e., those working at tertiary level or in a TESOL context, but similar findings are evident in the limited number of studies carried out with teachers for school contexts, as indicated below.

Breyer (2009) reported on pre-service secondary school language teachers in Germany who, after corpus training, were asked to design language-focused corpus tasks. A number of difficulties were reported, relating to choice of corpus, the tendency to create 'closed' tasks (where

there is clearly a right answer, thus circumventing the idea of discovery learning), assessing the difficulty level of their tasks for students, and technical problems. This led Breyer (2009) to call for ready-made and integrated tasks to be made available. Similarly, Shaeffer-Lacroix (2019) found that her trainee secondary school teachers of L2 German in France had difficulty when, after initial training, they were asked to design a learning task to include at least one corpus-based learning activity. Although the majority of them concluded that DDL was useful, only three of the ten participants said they would consider using corpora with younger learners. The need for sufficient training time, pedagogically relevant resources for DDL and user-friendly corpus analysis tools is emphasized as a result. Providing teaching activities for modelling purposes is also suggested, as is increasing familiarity with different approaches to reading, e.g., scanning vertically rather than reading horizontally, which is typical of most corpus-based activities.

Learner corpora have also been used in teacher training. Callies (2019) describes the use of a learner corpus of writing from younger learners (primary and secondary school pupils) when training pre-service teachers of English at German primary and secondary schools, on the basis that this more closely reflects the level the trainee teachers will work with in future. In a short training session, trainee teachers were guided to compare the use of intensifying adverbs in graded corpora, and at different grade levels in the learner corpus. They worked through a range of hands-on awareness-raising activities, encouraging them to explore language usage at different proficiency levels, to use corpora for evaluation purposes, and to consider how to develop pedagogical activities in their own classrooms. He concludes with a call for applied linguists to develop and implement modules on (learner) corpus linguistics for teacher students 'focusing on pedagogic aspects to empower teachers to actually teach with corpora' (Callies 2019: 261).

2.3 Online DDL teacher training

There has been relatively little empirical research into online teacher training in corpus linguistics. One study of an online introduction to corpora course with pre-service teachers concludes that trainees 'should be constantly given examples of real-life applications of the tools; merely working with different interesting CL [corpus linguistics] tools is of little value if there is no effort to link it to practice' (Ebrahimi and Faghih 2017:

129). The researchers also identified a need for a significant level of support from the instructors. This contrasts with Boulton's (2010) account of an online course in corpus linguistics as part of an MA in English, where students took a discovery approach to corpus linguistics after an initial introduction and worked quite autonomously to produce a 20-page report on a research project they defined and conducted themselves (see also Smitterberg, this volume). Boulton (2010), while acknowledging that support should be available, found that most students were generally able to work autonomously in a short period of time. Although this divergence may reflect different levels of digital, or other, confidence and expertise of the participants, it perhaps underlines the need for greater support when introducing corpus linguistics to be applied in a pedagogical context, as Ebrahim and Faghih (2017) were working specifically with teachers, while Boulton's (2010) students had wider-ranging interests.

2.4 Implications from previous research

In summary, then, previous research highlights the importance of teacher mediation of DDL to a given teaching context. This means that teachers need support in developing technical and corpus skills, gaining an awareness of what is available in a rapidly-evolving field, and, most importantly, identifying appropriate tasks for their students. They will need to make decisions about using pre-prepared materials, online or offline concordancing, corpus modality and the corpus interface (direct or indirect). It is therefore critical that their training has a clear focus on practical pedagogical applications, ideally from the outset (Ebrahim and Faghih 2017). Such training may be pre-service or in-service; in both cases ideally with continuing professional development in this area. Time is an essential requirement, both in terms of training time for teachers, and classroom time for DDL applications. The training demands are therefore to find time-efficient procedures to develop teachers' DDL skills, foregrounding potential DDL classroom applications while encouraging an understanding of the underlying theoretical principles. The remainder of this paper describes a pilot study introducing DDL into a teacher training course, with the principles above in mind. The outcomes of this are discussed, and approaches to further integrating DDL into the training programme are considered.

3. The study

3.1 Context

The study took place at a university in Sweden. Sweden seems ideally placed to include the integration of DDL into the secondary school classroom for several reasons. First, there has been a great deal of emphasis on the use of digital tools in schools. For example, it is recommended that all pupils from first grade upwards are provided with a personal laptop or tablet (Skolverket 2016), and the syllabus makes direct reference to the use of ICT, as discussed below. However, the adequacy of training in and information about pedagogical applications of ICT for teachers in Sweden has been questioned (e.g., Santos Muñoz and White 2020; Bunting et al. 2021). Bunting et al. (2021: 7), for example, reported that their group of 11 teachers (of English 4–6) 'were expected to find and choose proper digital tools on their own and preferably for free', relying on colleagues, social media, or even students for ideas. A second argument for promoting DDL regards English proficiency, which is generally high in Sweden[1] so the issues around DDL use with low-level learners (e.g., Moon and Oh 2018; Szudarski 2019) are less relevant. Finally, structures for professional development are in place. Skolverket (the Swedish National Agency for Education) facilitates qualified in-service teachers in gaining additional qualifications, including English, through *Lärarlyftet* programmes. Many of these programmes are provided as distance courses, as was the case in the present study.

The STs participating in this course were preparing to teach English to pupils in years 7–9 of compulsory school (13 to 16-year-olds), and to provide context for this, the curriculum is considered here. The curriculum for English is framed in communicative terms, with the general expectations for year 9 outlined in terms of receptive, productive and interactive ability. Those applicable to the present study are shown below (the *minimum*, grade E, requirements are shown, with the bolded phrases adapted in the descriptors for grades C and A):

Receptive ability: 'Pupils can understand **the main content and clear details** in English spoken at a moderate pace and in basic texts in various genres.'

Productive ability: 'In oral and written production in various genres, pupils can express themselves **simply, understandably and relatively**

[1] Sweden ranked fourth in the English First English Proficiency Index 2020.

coherently. To clarify and vary their communication, pupils can work on and make **simple** improvements to their communications.'
Interactive ability: 'In oral and written interaction in different contexts, pupils can express themselves **simply and understandably** and also **to some extent** adapted to purpose, recipient and situation. In addition, pupils can choose and apply **basically functional** strategies which **to some extent solve problems** and improve their interaction.' (Skolverket 2018: 39–40).

The curriculum closely follows the approach of the Common European Framework of Reference (CEFR) (Council of Europe 2001), and suggests the CEFR level associated with the different years of study (Skolverket 2011), with year 9 corresponding to level B1.1. This means, in terms of grammatical accuracy, that the student at this level '[u]ses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations' (Council of Europe 2001: 114). This ties in well with a DDL-based approach, which can help to expose lexico-grammatical patterning. Familiarity with DDL and corpus-based software also contributes to another core requirement of the curriculum, that of strengthening digital skills. An overall goal is that by the end of compulsory school, students 'can use both digital and other tools and media for attaining knowledge, processing information, problem-solving, creation, communication and learning' (Skolverket 2018: 12).

3.2 Course structure and participants

The course the STs were taking was an online six-module (45-credit) course taught over three semesters at half-pace. It was hosted in Moodle, where all course materials (timetables, assignments, recorded lectures, readings, links etc.) were stored and discussion forums could be accessed. Discussion forums, whether written or oral (via VoiceThread) were asynchronous to facilitate access outside working hours, supplemented by optional Zoom tutorials. This approach to English courses had already been established at the university, pre-Covid 19 pandemic, to allow access to education to a wide range of students living and working throughout the country and abroad.

The STs were 13 in-service secondary school teachers (nine female and four male), most working full-time while studying. This was the first 7.5 credit module, entitled 'Grammar and English language teaching'. It was divided into five sub-modules, each focusing on a grammar topic and

associated teaching approaches. The STs taking this course came from a wide range of disciplines, e.g., maths, physics, religion, music, art, and had diverse teaching experience. Some had been teaching English at lower levels, and/or other languages such as Swedish and Spanish. All of them were attending the course to upskill and gain the required formal qualification to teach English for years 7–9. Two STs were native English speakers, while English was a second or third language for others, with Swedish, French and Arabic as first languages. All met the standard requirement for studying English at university level, English 6, corresponding to the international qualifications Cambridge Certificate in Advanced English (CAE) or an IELTS score of 6.5. All of the STs were new to the university and to online study.

The sub-module described here was called *Lexical grammar*, and it was the third two-week sub-module to be studied. It introduced STs to an alternative approach to language analysis, using DDL in order to examine lexico-grammatical patterning. The other sub-modules in the course take a more traditional approach, and for these STs, this was their first encounter with corpora and corpus software.

3.3 DDL interface

Boulton (2010) suggests a number of guiding principles to facilitate the introduction of DDL; that tools should be free, the software should be easy to use, and the techniques should be flexible and transferable. These informed the methodological decisions taken. The interface adopted was Voyant Tools (Sinclair and Rockwell 2016). Although not widely used in DDL studies, it has a number of advantages, being freely available, web-based and stable, with a simple and relatively intuitive interface. It is also one of the few online corpus applications into which independent corpora can be easily uploaded, as many web concordancers are tied to a dedicated corpus, e.g., Sketch Engine for Language Learning (SKELL) (Baisa and Suchomel 2014), or a range of inbuilt corpora, e.g., Lextutor (Cobb 2021). Figure 1 shows the standard interface that appears on Voyant Tools once the corpus has been uploaded. The corpus text is shown centrally, and any word in it can be highlighted by clicking on it, which then brings up concordance lines of the word in the bottom right-hand box (which can be expanded to show the full context with a click). The word cloud on the upper left visually indicates the most frequently-used words (giving the raw frequency of the word when the cursor is hovered over it). Details of

the corpus are given on the bottom left (total word count, number of types, vocabulary density, average words per sentence, and most frequent words in the corpus), and the location and occurrences of the most frequent words in the corpus are shown in the chart on the top right. The page is customisable, with other features interchangeable by clicking on the icons at the top of each box, e.g., the summary box on the bottom left gives a useful list of most frequent phrases when this option is selected. Explanations are given for all of the features when the cursor is hovered over the word.

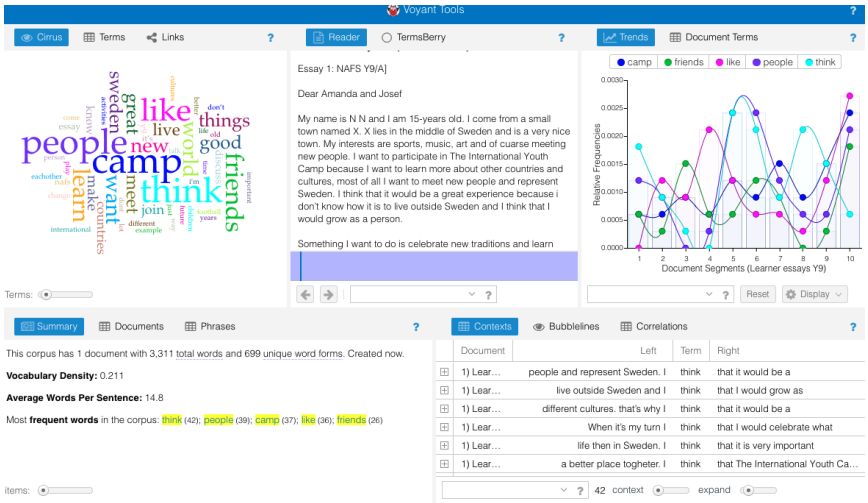

Figure 1: Standard interface of Voyant Tools (Sinclair and Rockwell 2016), shown with the learner corpus

3.4 Corpora

The Voyant Tools interface opens onto an 'Add texts' page where the selected corpus is uploaded. There is an option to type in URLs or paste in texts in the window, to choose one of two pre-loaded English literature-based corpora, or to upload files. For the present study, two corpus files were compiled for the STs to use: a learner corpus consisting of a set of learner essays, and a graded corpus of published texts appropriate to the learners' level. In each case, the corpus needed to be large enough to demonstrate repeated patterns, but since they consisted of texts that were

limited in terms of word type frequency, they could be reduced in size and still show some results (see McCarthy and Carter 2001).

The learner corpus was compiled from publicly-available writing samples on the National Tests in Foreign Languages Project (NAFS) website (University of Gothenburg). These were used to ensure that the level was appropriate, and that the language range was narrow by limiting it to one topic. It consisted of nine essays, three of each graded at levels A, C and E according to the National Test grading criteria, making a small corpus of 3,300 words. Although this was very limited, as STs might expect to work with a similarly-sized corpus made up of work from their own classes, it was useful to see if this could be used to produce meaningful results. For the graded corpus, 10,000 words of reading texts at the pupils' English level were collected, including texts (letters, articles, stories) from the NAFS website, English language teaching websites, and coursebooks targeting years 7–9. The intention was to provide examples of language that pupils at this level would be likely to meet. For practical purposes, each corpus was presented as a single document, with each text given an identifier, rather than multiple individual files. This meant that the STs were only required to upload one document, either the learner or the graded corpus, into Voyant Tools at any one time, reducing the potential for confusion or error.

3.5 Procedure

On Moodle, the STs were first provided with an introduction to corpus linguistics and its applications within the classroom. Texts consisted of selected chapters that gave an accessible and practical overview to the use of corpora in English language teaching (Bennett 2010; Frankenberg-Garcia 2016). Audio-visual material included a pre-recorded lecture on lexical grammar and related links to web-based presentations by academics and teacher trainers. A dedicated section on teaching applications was included, with further links to short articles and ideas. Finally, a resources section provided links to the Voyant Tools web-based concordancer, a screencast demonstrating how to use the application, and the corpus files described above.

The STs were given two short practical tasks to complete, driving them to gain some hands-on experience. In the first task, they needed to use Voyant Tools with the learner corpus to find examples of good usage and errors in phraseology (i.e., collocation/idiomatic use) for two words.

They were specifically asked not to focus on spelling errors. An example was given as shown in Table 1. Note that for this example, a word with few hits was selected, both to avoid using the more frequent examples, and also to demonstrate that they did not need to find a word with many hits.

Table 1: Example given on learner corpus task

Word 1: Big	Total hits*: 2
Good example:	violence is a big issue
Error:	racism wouldn't be as big as it is now

*number of times the word occurs

The second task directed them to use the graded corpus to find two collocations or idioms that occurred at least twice, which they thought would be useful to teach to their students. Again, an example was provided, as shown in Table 2.

Table 2: Example given on graded corpus task

Collocation/idiom 1: the vast majority of	Total hits: 2
Context 1: Studies suggest that the vast majority of teenagers do not really worship celebrities.	
Context 2: However, opponents of zoos say that the vast majority of captive breeding programmes do not release animals back into the wild.	

In order to complete these tasks, the STs needed to refer to the screencast. They had the opportunity to attend a live Zoom seminar if they wanted further guidance, and could discuss questions on the forum online. Two students attended the live seminar, and three asked for clarification on the forum (which all students could view).

Once they had completed these tasks, the STs were required to contribute to a VoiceThread assignment, presenting a short recording of their thoughts in response to the following question, taken from one of their initial readings (Frankenberg-Garcia 2016): 'In your own professional context, to what extent is it realistic to expect teachers to develop corpus skills and use corpora with learners?' They were asked to motivate their responses, considering issues like time, technology, interest, learning styles etc.

The task responses were submitted to Moodle; each student received individual feedback on their work, and a summary of the examples found for each task was compiled and shared on Moodle. The tasks were also analysed within the context of the pilot, as discussed in section 3.6. Similarly, the VoiceThread recordings were transcribed, analysed manually, and a brief summary of the themes that had emerged was posted on Moodle, with suggestions for ways forward. A more detailed account of the thematic analysis is given in section 3.7.

3.6 Task responses

As all of the STs were unfamiliar with using DDL, some found the assignment difficult initially and lacked confidence in their ability, but all of them were able to complete the two practical tasks. For the first task, the majority of the STs found appropriate examples, as in Table 3, where the ST has identified a good use and an error using *learn*. Other responses included verbs such as *make*, *told*, and *like*, collocations with certain words, such as *great* and a number of grammatical function words, such as *that*, *the*, *to*, and *of*. However, two unclear/inappropriate examples were given from two different STs, one of which is shown in Table 4. It may be that these examples have simply been misallocated, suggesting that a useful extension to the task would be to ask the STs to explain *why* they considered their examples incorrect, for example, or *how* they might correct them.

Table 3: Student response to Learner corpus task, considered correct

Word 2: Learn	Total hits: 17
Good example:	we'll learn a lot about each other
Error:	Then they'll learn their children to accept everybody

Table 4: Student response to Learner corpus task, considered incorrect

Word 2: People	Total hits: 30
Good example:	among the people who lives in it
Error:	talk to the people from other contries

The second part of the task proved to be easier for the STs, with all of them able to identify frequent phrases that would be useful to teach to their students. These included phrasal verbs, (e.g., *took off*) linking expressions

(e.g., *by means of, on the other hand, as well as, as part of, all the way to*), quantifiers (e.g. *a number of*), sentence stems (e.g. *are more likely to*), and collocations (e.g. *digital footprint, family home, out of control*).

3.7 Participants' feedback

Feedback was collected from the VoiceThread assignment, where the STs reflected on the potential for implementing what they had learned in a classroom context (see section 3.5). All students participated in this. Although the STs' comments were largely positive, they expressed a number of reservations. They mentioned the value of DDL for checking language knowledge, for finding authentic examples to demonstrate use, for creating learning materials, e.g., gapfill exercises, and for focusing on phrasal language. Some felt that using Voyant Tools would appeal particularly to some students: 'Some students might find it exciting to find their own verification' according to one ST, another mentioned encouraging more autonomy, while another pointed out the potential for differentiation, for higher-level students in a group to engage with DDL. However, they unanimously agreed that DDL should be used at the discretion of the teacher, following 'adequate training'.

Time constraints were mentioned by all of the STs, with comments such as 'I'm intrigued by the possibilities but time is a barrier'. This included time to gain confidence in their own competence in the use of corpus tools; one ST said, 'I'm not confident enough to use this in class but plan to develop my own skills', while another felt that teachers should not attempt to use corpora 'unless they have had adequate training'. The need to work out how it could be integrated into the syllabus and with their students was also mentioned, one ST saying they 'need more clarity on the purpose and benefits'. Given that students then also need to be trained, some STs felt that this would make it unfeasible in some teaching situations, one stating '[i]t's not realistic—teachers have so little time'.

Technology was another common theme. Internet and computer access was not considered an issue, but there were some concerns about learning to navigate the software. Some STs indicated that they (and their colleagues) feel overwhelmed with the tools they already have; some teachers are not interested in finding out about more, or they may find them difficult to use without training. There were contradictory views on the accessibility of Voyant Tools, two STs saying that they found it intuitive, e.g., 'I was surprised at how user-friendly Voyant Tools was',

while two others said they found it difficult to navigate, e.g., '[t]he software wasn't so user-friendly'. One ST pointed out that although we assume younger people to be computer literate, the type of intuitive applications they commonly access means they have little ability to navigate computer use beyond this, and they may find the Voyant Tools interface off-putting.

In terms of planned uptake following this initial task, three of the 13 STs expressed enthusiasm about DDL, and four planned to explore it further and try it out, hands-on, in class. Another ST reported being 'really interested to use this with my students' essays' as an analytical tool. Several STs suggested that if they used it, they would begin with a hands-off approach, using pre-prepared exercises, and follow up with students who may be interested in using online DDL. Two more STs felt that they would be prepared to use it if they had more specific training in the purpose and benefits of using DDL. Another ST suggested that simpler tools, like the word cloud shown in Voyant Tools would be more interesting to her students than concordancing, and she would be more likely to explore this with them. Two STs felt that this approach would not necessarily suit the average learner in 7–9, one suggesting that it would be more suitable for upper-secondary school students, and the other that it would be better for university-level students.

3.8 Reflection on the pilot study

The pilot was considered to provide an effective introduction to DDL; STs engaged positively with the tasks, and these were achievable with the level of scaffolding provided. Voyant Tools as an interface had a mixed reaction, but a benefit was that it introduced them to a range of tools rather than only a concordancer, allowing STs see more visual applications (e.g., word cloud) which one ST saw potential for with her students. In terms of corpora, the STs found it easier to work with the larger graded corpus than the learner one. This may have been due to its very limited size, or to a lack of confidence in their own English language competence and ability to identify mistakes. Some STs intuitively approached the learner corpus as an essay set, i.e., reading through the essays for errors. As DDL encourages both horizontal and vertical reading, finding errors through conventional reading in this way then using corpus tools to find the frequency of such errors is a good approach. However, it would have been more effective with a larger learner corpus for STs to get a better sense of

common error types. The size of the graded corpus worked well; STs found that it contained sufficient examples of useful language but was limited enough not to be overwhelming. The tasks were achievable, although they could be improved by including a peer-reviewing stage to encourage the STs to articulate their procedures and justify their outcomes, thereby avoiding some of the ambiguities that emerged.

Although the overall reaction was positive and STs could see how DDL could be helpful, they expressed reservations about using DDL in the classroom. An introduction that was more clearly grounded in classroom practice may have produced a more enthusiastic response. Further extension and consolidation of their knowledge is also needed. On this programme, STs work through a range of modules, and DDL could be integrated into some or all of them with the aim of normalising it as a classroom activity. Some suggestions for how to do this are discussed below.

4. An integrated approach to DDL training

Drawing on research into the effective use of digital applications in general in the English classroom, the value of *gamification*, i.e., applying game-like elements to a non-game context (Werbach and Hunter 2012) has been noted (see Figueroa 2015 for an overview). One example noted in Bunting et al.'s (2021) study is the Google *QuickDraw* application[2]. Applications such as *Kahoot* and *Quizlet* are also popular, with reputable websites such as Cambridge English providing kahoots for younger learners[3]. Further research in this area is required, but it seems that digital tools are more likely to be adopted when they are freely accessible, require little preparation, and have an interactive and/or competitive element which makes them motivating and fun for the students. If we apply this to DDL, Meunier's (2020: 18–19) proposal of the PlayPhrase.me app or Web interface as a pedagogical tool is a good starting point (see 2.1). As Meunier (2020: 19) suggests, this could be used for competitive activities such as 'Find four expressions using *give me a*'. Many of the types of discovery activities used with DDL in text could be used, such as finding

[2] https://quickdraw.withgoogle.com/

[3] https://www.cambridgeenglish.org/teaching-english/resources-for-teachers/kahoot/

different meanings and uses of phrasal verbs, e.g., *take off*, exploring differences in use in different moods and tenses (e.g., *take off* vs. *took off*). It is important to note that the video clips are not censored in any way and may include language inappropriate for the classroom, and there are occasional inaccuracies in the subtitles (*intact* for *attacked*, for example). However, as Meunier points out, the app is quite addictive and offers a novel way of approaching DDL, with added phonological benefits. Meunier's other suggestions (see 2.1) offer similar, motivating ways of using apps for pedagogical purposes. These may offer a better entry point into DDL than the tasks used in the present study, as the STs can immediately see their classroom potential.

Shortages of time and training have been consistently noted as barriers to teachers' use of DDL, both in the present and previous studies. Pre-prepared materials may help overcome this, either used as presented or as a model for developing customised materials (Breyer 2008; Shaeffer-Lacroix 2019). Introducing STs to the worksheets available on Lancaster University's Corpus for Schools website[4] might be helpful in this regard. This site provides ELT handouts that explore aspects of current language use, spoken communication and communicative skills, drawing on the British National Corpus 2014 and giving an introduction to the BNClab interface (Brezina et al. 2018). Engaging sociolinguistic topics are covered, such as why people use swearwords in English, and how people talk about Christmas. Although they may be pitched at too high a level for 7–9 students, they provide an opportunity for STs to develop their own skills in corpus use, as well as being a model for materials development. Pre-prepared materials using learner corpora could be another starting point. For example, speaking activity worksheets[5] based on the Trinity Lancaster Corpus (Gablasova, Brezina and McEnery 2019) explore features of successful communication, based on speaking tests from learners at B1–C2 level. Another resource it would be valuable to introduce STs to is the Lextutor website[6], in particular the Cloze and Multi-Concordance tools. Once STs become familiar with these, they provide time-efficient way of developing corpus-based tasks that they can tailor to their students' needs.

[4] http://wp.lancs.ac.uk/corpusforschools/esl-teaching-materials/
[5] https://www.trinitycollege.com/about-us/research/Trinity-corpus/corpus-resources/classroom-activities
[6] https://www.lextutor.ca/

Using DDL in exploring English literature presents another opportunity, as literary texts in digital form can be similarly analyzed using corpus tools. One application STs could be introduced to is the CLiC project[7] (Mahlberg et al. 2016), which provides a web-based concordance linked to a range of texts, including a corpus of 65 children's books, which can be studied as a whole or as individual books. The CLiC activity book (Mahlberg et al. 2017) provides training and suggestions in using corpus tools for literary analysis. Although many activities may be too in-depth for the 7–9 level, one accessible approach is to use a literary text to create a word cloud of the most frequently occurring words, indicating characters and themes. Using literary dialogue as a model of conversation for English language learners is another approach, and an adaptable worksheet is provided for this (Jones and Oakey 2019).[8]

In the process of the course, STs could also be introduced to free user-friendly applications for more traditional approaches to DDL. SKELL (Baisa and Suchomel 2014), for example, offers a clean interface with three useful options for the target word: a list of examples, a word sketch showing common collocations, and similar words. Although this deviates from traditional DDL in that the example sentences cannot be manipulated, the word sketch extracts the relevant information, i.e., differentiates between a noun as subject and object, indicates frequently used adjectives and modifiers, phrasal expressions etc., and allows the user to click on the collocates to see further examples of the words used together. *Just-the-word* is another free and popular tool that gives a quick picture of most-used collocations drawn from the BNC. A more advanced tool is Collocaid (Frankenberg-Garcia et al. 2019) which offers suggestions for collocations when writing texts. Although currently only available as a prototype[9], as it evolves it is likely to be a useful tool for English language learners at all levels.

5. Conclusion

This paper describes a pilot study introducing DDL to a group of STs in Sweden, informed by previous research in this field. Although successful

[7] https://clic.bham.ac.uk/

[8] Available at https://www.birmingham.ac.uk/Documents/college-artslaw/elal/clic/Jones-Oakley-2019-CLiC-dialogues-activity.pdf

[9] https://www.collocaid.uk/

at a superficial level, the sub-module described is far from sufficient to encourage teachers to use it as a classroom tool. This reflects the findings of other studies with school-age students. As noted in the introduction, although many researchers and tertiary-level language learners see the value of DDL, it has not been adopted within language learning at a broader level. This is perhaps not surprising, given that the traditional approach requires the analytical processing of a great deal of written text. However, DDL can now be conceived of in different ways, using a range of resources. Novel approaches using multimedia, more streamlined interfaces and high-quality pre-prepared materials are all available, and these can be tailored to younger learners and integrated into teacher training courses, giving STs a broader view of DDL and its various pedagogical applications. If teachers are shown easily-prepared tasks with appeal for their learners, they are more likely to add them to their teaching repertoire.

There is no doubt, however, that gaining traction for DDL in the secondary classroom is a challenge. Even in the Swedish system, training opportunities for teachers are limited, making it both difficult to generate sufficient awareness of DDL, and to keep pace with new applications as they emerge. From the students' perspective, young learners have increasingly high expectations, digitally speaking, and the interfaces available may fall short of these. Further research into approaches to training pre-tertiary-level teachers in DDL is needed to help to develop more effective practices, as are studies into teachers' and students' responses to newer pedagogical applications using corpus-based techniques. As Boulton and Vyatkina (2021) point out, we need to find out more about what works in different contexts and for different learner profiles. It is hoped that such studies can feed into the development of further training packages, tools and approaches, as we continue to expand our view of how corpus data can inform teaching and learning.

References

Allan Rachel. 2009. Can a graded reader corpus provide 'authentic' input? *ELT Journal* 63: 23–32. doi:10.1093/elt/ccn011.

Baisa, Vít, and Vít Suchomel. 2014. SkELL – Web interface for English language learning. In *Proceedings of recent advances in Slavonic natural language processing RASLAN 2014*, edited by Aleš Horák and Pavel Rychlý, 63–70. Brno: Tribun EU.

Bennett, Gina R. 2010. *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor: University of Michigan Press.

Boulton, Alex. 2010. Bringing corpora to the masses: Free and easy tools for interdisciplinary language studies. In *Selected papers from Teaching and Language Corpora 2006*, edited by Nathalie Kübler, 69–96. Frankfurt: Peter Lang.

Boulton, Alex. 2017. Corpora in language teaching and learning. *Language Teaching* 50: 483–506. doi:10.1017/S0261444817000167.

Boulton, Alex, and Nina Vyatkina. 2021. Thirty years of Data-Driven Learning: Taking stock and charting new directions over time. *Language Learning & Technology* 25(3): 66–89. doi:/10125/73450.

Breyer, Yvonne. 2009. Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning* 22: 153–72. doi:10.1080/09588220902778328.

Brezina, Vaclav, Dana Gablasova, and Susan Reichelt. 2018. BNClab. http://corpora.lancs.ac.uk/bnclab [electronic resource], Lancaster University.

Bunting, Leona, Ylva Hård af Segerstad, and Wolmet Barendregt. 2021. Swedish teachers' views on the use of personalised learning technologies for teaching children reading in the English classroom. *International Journal of Child-Computer Interaction* 27: doi:10.1016/j.ijcci.2020.100236.

Callies, Marcus. 2019. Integrating corpus literacy into language teacher education: The case of learner corpora. In *Learner corpora and language teaching*, edited by Sandra Götz and Joybrato Mukherjee, 245–263. Amsterdam: Benjamins.

Chambers, Angela. 2019. Towards the corpus revolution: Bridging the research-practice gap. *Language Teaching* 52: 460–475. doi:10.1017/S0261444819000089.

Cobb, Tom. 2021. *Compleat Lexical Tutor* v.8.3. Accessed September 2021. https://www.lextutor.ca/.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Crosthwaite, Peter (ed.). 2019. *Data-Driven Learning for the next generation: Corpora and DDL for pre-tertiary learners*. London: Routledge.

Ebrahimi, Alice, and Esmail Faghih. 2017. Integrating corpus linguistics into online language teacher education programs. *ReCALL* 29: 120–135. doi:10.1017/S0958344016000070.

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

Figueroa, Jorge. 2015. Using gamification to enhance second language learning. *Digital Education Review* 21: 32–54.

Frankenberg-Garcia, Ana. 2012. Raising teachers' awareness of corpora. *Language Teaching* 45: 475–489. doi:10.1017/S0261444810000480.

Frankenberg-Garcia, Ana. 2016. Corpora in ELT. In *The Routledge handbook of English language teaching*, edited by Graham Hall, 383–398. London: Routledge.

Frankenberg-Garcia, Ana, Robert Lew, Jonathan Roberts, Geraint Rees, and Nirwan Sharma. 2019. Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL* 32: 23–39. doi:10.1017/S0958344018000150.

Gilquin, Gaëtanelle, and Sylviane Granger. 2010. How can Data-Driven Learning be used in language teaching? In *The Routledge handbook of corpus linguistics*, edited by Anne O'Keeffe and Michael McCarthy, 359–370. London: Routledge.

Jones, Christian, and David Oakey. 2019. Literary dialogues as models of conversation in English language teaching. *Journal of Second Language Teaching and Research* 7: 108-135.

Kaltenböck, Gunther, and Barbara Mehlmauer-Larcher. 2005. Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL* 17: 65–84. doi:10.1017/S0958344005000613.

Karras, Jacob Nolen. 2016. The effects of Data-Driven Learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL* 28: 166–186. doi:10.1017/S0958344015000154.

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

Leńko-Szymańska, Agnieszka. 2017. Training teachers in Data-Driven Learning: Tackling the challenge. *Language Learning & Technology* 21: 217–241.

Mahlberg, Michaela, Peter Stockwell, Johan de Joode, Catherine Smith, and Matthew Brook O'Donnell. 2016. CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora* 11: 433–463.

Mahlberg, Michaela, Peter Stockwell, and Viola Wiegand. 2017. *CLiC – Corpus linguistics in context: An activity book.* Available at https://www.birmingham.ac.uk/Documents/college-artslaw/elal/clic/CLiC-Activity-Book-license.pdf.

McCarthy, Michael, and Ronald Carter. 2001. Size isn't everything: spoken English, corpus, and the classroom. *TESOL Quarterly* 35: 337–340. doi:10.2307/3587654.

Meunier, Fanny. 2019. A case for constructive alignment in DDL: Rethinking outcomes, practices and assessment in (data-driven) language learning. In *Data-Driven Learning for the next generation: Corpora and DDL for pre-tertiary learners*, edited by Peter Crosthwaite, 13–30. London: Routledge.

Meunier, Fanny. 2020. Revamping Data-Driven Learning: Making actual use of the affordances of digital technology. Paper presented at the online symposium *Incorporating Corpora in Teaching*, Mid-Sweden University, Sundsvall, Sweden, 23 October.

Mirzaei, Azizullah, Masoud Rahimi Domakani, and Sedigheh Rahimi. 2015. Computerized lexis-based instruction in EFL classrooms: Using multi-purpose LexisBOARD to teach L2 vocabulary. *ReCALL* 28: 22–43. doi:10.1017/S0958344015000129.

Moon, Soyeon, and Sun Young Oh. 2018. Unlearning overgenerated *be* through Data-Driven Learning in the secondary EFL classroom. *ReCALL* 30(1): 48–67. doi:10.1017/S0958344017000246.

Naismith, Ben. 2017. Integrating corpus tools on intensive CELTA courses. *ELT Journal* 71: 273–283. doi:10.1093/elt/ccw076.

O'Keeffe, Anne, Michael McCarthy, and Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Pérez-Paredes, Pascual. 2019. The pedagogic advantage of teenage corpora for secondary school learners. In *Data-Driven Learning for the next generation: Corpora and DDL for pre-tertiary learners*, edited by Peter Crosthwaite, 67–87. London: Routledge.

Santos Muñoz, Arantxa, and Jonathan White. 2020. Digitalisation in schools in Sweden: Needs and experiences of trainees of languages. *Lingua* 2: 28–31.

Shaeffer-Lacroix, Eva. 2019. Barriers to trainee teachers' corpus use. In *Data-Driven Learning for the next generation: Corpora and DDL for Pre-tertiary Learners*, edited by Peter Crosthwaite, 47–64. London: Routledge.

Sinclair, Stéfan and Geoffrey Rockwell, 2016. *Voyant Tools*. Accessed September 2021. http://voyant-tools.org/.

Skolverket. 2011. Om ämnet Engelska. Accessed August 2021. https://www.skolverket.se/download/18.6011fe501629fd150a28916/1536831518394/Kommentarmaterial_gymnasieskolan_engelska.pdf.

Skolverket. 2016. IT-användning och IT-kompetens i skolan. Rapport 2015:00067. Accessed September 2021. https://www.skolverket.se/download/18.6bfaca41169863e6a65c32a/1553966893852/pdf3667.pdf.

Skolverket. 2018. Curriculum for the compulsory school, preschool class and school-age educare 2011 (Revised 2018). Accessed September 2021. https://www.skolverket.se/download/18.31c292d516e7445866a218f/1576654682907/pdf3984.pdf.

Soruç, Adem, and Bilal Tekin. 2017. Vocabulary learning through Data-Driven Learning in an English as a second language setting. *Kuram ve Uygulamada Egitim Bilimleri* 17: 1811–1832. doi:10.12738/estp.2017.6.0305.

Szudarski, Pawel. 2019. Effects of Data-Driven Learning on enhancing the phraseological knowledge of secondary school learners of L2 English. In *Data-Driven Learning for the next generation: Corpora and DDL for Pre-tertiary Learners*, edited by Peter Crosthwaite, 133–149. London: Routledge.

University of Gothenburg. 2020. National tests of foreign languages project. Accessed September 2021. https://www.gu.se/en/national-tests-of-foreign-languages.

Werbach, Kevin, and Dan Hunter. 2012. *For the win: How game thinking can revolutionize your business*. Philadelphia: Wharton Digital Press.

Wicher, Oliver. 2019. Data-Driven Learning in the secondary classroom: A critical evaluation from the perspective of foreign language didactics. In *Data-Driven Learning for the next generation: Corpora*

and DDL for Pre-tertiary Learners, edited by Peter Crosthwaite, 31–46. London: Routledge.

Zareva, Alla. 2017. Incorporating corpus literacy skills into TESOL teacher training. *ELT Journal* 71: 69–79. doi:10.1093/elt/ccw045.