LANGUAGE CORPORA AND DATA-DRIVEN LEARNING IN SECOND LANGUAGE ACQUISITION

Lilit Avetisyan
PhD in Applied Linguistics ANPP Training Center
lilit.avetisyan@mic.ul.ie
Marine Avetisyan
EUA, Chair of Languages
maravetis@yahoo.com

Abstract

The present article focuses on the need to address language corpora and data-driven learning (DDL) as one technology-based approach to language learning that can bring the real language use into the classroom, offer new tools and support learning, and expand opportunities for self-directed learning. However, this potential of language corpora is still not a mainstream methodology in second language acquisition. The article discusses the pedagogical context of DDL, underpinned by the theory of constructivism, and presents the direct, computer-based and indirect, hands-off approaches of DDL. Additionally, it argues for the direct engagement with language corpora if our aim is to achieve such long-term benefits as attention, awareness, and autonomy. The study also brings to light some of the fears, challenges, and benefits of using DDL to mitigate the risks of the uncritical use of corpus tools in the language classroom and enhance their impact on the efficiency of language learning.

Keywords: language corpora, second language acquisition, data-driven learning, hands-on and hands-off DDL.

Introduction

Throughout the history of second or foreign language teaching, over 60 theories, models, hypotheses, and perspectives have been proposed. However, there is no consensus regarding the effectiveness of a particular teaching approach in accelerating the acquisition of and facilitating the automatization of taught knowledge. Thus, the **aim** of the present study is to analyze a technology-based approach, namely data-driven learning (DDL), which is believed to provide corpus-based solutions to some of the concerns in language pedagogy. The research **objectives** set are as follows: to explore the theoretical underpinnings of DDL, to identify the benefits and challenges of different approaches of DDL, to provide an account and evaluation of the current pedagogical context of DDL. Descriptive and synthesis **methods** have been applied to investigate the literature dominating the field under study and the current issues regarding the implementation of DDL in language learning practices.

In the world of corpus linguistics, a corpus is a large, principled collection of naturally occurring texts stored electronically. Corpus linguistics equips teachers and learners with confidence that they are learning the language they will encounter outside the language classroom and in the real world of language use [Reppen, 13-21]. Corpora are "records of language behavior" which represent a wealth of knowledge about language [Cook, 57-64]. They provide knowledge of "linguistic and co-occurrence patterns", which would be difficult to otherwise identify [Reppen, 13-21]. EFL/ESL professionals repeatedly make decisions about language and the choice of lexico-grammatical features to teach and to test. They also attempt to use authentic rather than made-up learning materials. "Invented examples can present a distorted version of typicality or an over-tidy picture of the system" [Kennedy, 318]. "Corpora have also brought to light features about language which had eluded our intuition" [O'Keeffe, 21]. They provide knowledge about what have actually been said, not what can be said. Thus,

corpora, as repositories of authentic texts, can serve as a source of descriptive insights for language teaching/learning and used as tools that directly influence the teaching/learning process [Bernardini, 165-182].

Language Corpora in Second Language Pedagogy

The development of corpus linguistics in recent years has highlighted the potential of corpora for language pedagogy [Johns, *DDL Examples,* 1-16). Several general corpora are readily available, including Brown; Lancaster, Oslo, Bergen corpus (LOB); British National Corpus (BNC); the corpus of Contemporary American English (COCA), the International Corpus of English (ICE), and provide valuable resources for information on how spoken and written language are used in a range of settings [Reppen, 13-21].

Language corpora are virtually used in the construction of reference materials, such as dictionaries (e.g. Rundell's Macmillan English Dictionary), grammar books (e.g. Biber, Johansson, Leech, Conrad, and Finegan's Longman Grammar of Spoken and Written English), usage manual (e.g. Swan's Practical English Usage), and textbooks (e.g. McCarthy, McCarter, and Sandiford's Touchstone). Corpora provide information on usage in the form of concordance with a key word highlighted in context (KWIC), on frequency, distribution, collocation, etc. Corpus linguistics also includes the use of language corpora, where learners are engaged in hands-on experience through guided activities or through corpus-based handouts with concordance lines. This experience relies on inductive approach, which enables learners to see the linguistic patterns of the target item and form generalizations [Johns, *On DDL Examples,* 1-16]. This type of learning is commonly referred to as 'data-driven learning' (DDL), which "confronts the learner as directly as possible with the data to make him/her a linguistic researcher" [Johns, *On DDL Challenge,* 108). The computer-based approach was coined by Tim Johns [*On Micro-Concord,* 151-162], who initially used corpora as a tool for language learners and contributed a lot of corpus-based teaching materials. DDL was defined as "the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language, and the development of activities and exercises based on concordance output" [Johns, *On DDL Examples,* 1-16].

Data-Driven Learning (DDL) and Constructivism

The pedagogical context of DDL fits well with the constructivist paradigm for language learning and the developments within the area of learner autonomy [Chambers, Kelly, 20-21]. DDL is based on Schmidt's [1-63] Noticing Hypothesis, according to which conscious attention is required for language learning to take place. In contrast to the "artificial" intellectual activity of trying to learn and use the rules, DDL allows learners to detect through their adaptive behavior language patterns that are meaningful to them, thus making learning more "natural" [Gaskel, Cobb, 304]. DDL provides authentic input for learners based on naturally occurring language. Kennedy [318] posits that "invented examples can present a distorted version of typicality or an over-tidy picture of the system". In line with this, as referred to above, O'Keeffe et al. (21) emphasize that "corpora have also brought to light features about language which had eluded our intuition". The appeal of corpora is that it introduces language not as what can be said but as what has actually been said.

DDL can potentially promote learners' active participation in the learning process by means of discovery of language rules by themselves based on their own exploration and analysis of concordance input. If learning is an act of discovery per se, learning takes place in a problem-solving environment, which requires learners to reason inductively – observe, classify, and generalize [Johns, On *DDL Examples,* 1-16]. Moreover, unlike the rule-based language learning, which separates grammar and lexis, DDL exposes learners to the target item as frequently occurring lexico-grammatical patterns [Flowerdew, 15-36]. It helps them to "identify linguistic and situational co-occurrence patterns", which are otherwise difficult to obtain [Reppen, 14]. This is also believed to facilitate the development of learner autonomy. By practicing noticing and raising consciousness, learners gain better learning skills, become more autonomous and better language learners outside the

classroom [Johns, *On DDL Examples,* 1-16]. O'Sullivan [277] provides an impressive list of cognitive skills that DDL can develop: "predicting, observing, noticing, thinking, reasoning, analyzing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorizing, hypothesizing, and verifying". Consideration of these advantages makes DDL an effective approach in second language learning.

However, like any other teaching/learning approach, DDL has not been left uncriticized. The reason that the concerns arise may be that the research so far has not succeeded in convincing wider audience about the payoffs of DDL approach in terms of the invested time, money, efforts, and resources [Boulton, *On Meta-Analysis,* 348-393]. This is due to the fact that most studies are either small-scale and qualitative or focus on learners' behavior working with a corpus, their attitudes towards DDL, and the use of a corpus as a reference tool. Even those studies that use quantitative design provide statistically non-significant results. Moreover, much empirical research is not concerned with long-term learner performance, finding it difficult to design; therefore, the focus falls on immediate learning outcomes. Most studies are conducted with advanced level learners, and only four with lower levels, which is partly responsive to the common belief that only advanced learners can benefit from DDL. However, there is also a belief that DDL can be no less useful for lower level students, and even more beneficial than for advanced learners [Boulton, *On Meta-Analysis,* 348-393]. Another concern is related to technological considerations. The reasons can be the lack of computers or insufficient technical backup, the considerable training required for effective DDL, or the irrelevance of DDL to local contexts, as perceived by teachers [Gabrielatos, 1-37].

Text-driven approach is a potentially effective way of exploiting experience of authentic texts. Since corpora contain authentic native language which is beyond the proficiency level of many learners, concern arises related to the authenticity of contrived language examples to which learners are exposed to. The aim of language teaching is to produce effective and competent communicators, which can be achieved by exposing them to input that exemplifies the real language; hence it should be authentic rather than contrived examples of data. Corpus, as a repository of authentic texts, can assist this aim. However, there is a view that numerous examples of texts in corpora, which were initially produced for a certain audience and not for a language learner, are decontextualized or taken away from their authentic context and reproduced in a teaching context, which may not meet the communicative goal of the classroom. Moreover, culturally-embedded texts might make it difficult for language learners to 'authenticate' the language for themselves. This suggests that authenticity should be defined as a relationship between a text and the response that it triggers in its immediate audience [Widdowson, 2-25]. Or as in Mishan's [346] refined definition of authenticity, learning tasks involving authentic materials should be correspondingly authentic, entailing interactions that are consistent with the original communicative purpose of the authentic text.

As a result, many practitioners support the use of contrived or culturally 'neutral' examples for pedagogical purposes. This way the learning input can be graded for different levels of language proficiency and be sensitive to learners' language learning needs [O'Keeffe et al., 314]. On the other hand, many others, who emphasize language authenticity, explain that the natural human predisposition will allow language learners to contextualize the authentic data for themselves and increase their motivation because the language they deal with is so 'real'. Still others recommend careful selection of materials from authentic sources which are easily contextualized, or tasks that can be graded to correspond to the nature of authentic data. O'Keeffe et al. [314] posit that teachers should use freely selected, carefully mediated, and locally relevant naturally-occurring examples rather than contrived or unreal examples – a responsibility that they have historically done and that is currently harnessed with technical possibilities of faster searches for authentic data.

In relation to this, it would be appropriate to discuss here the direct or computer-based and indirect or paper-based uses of data-driven learning.

Hands-on versus Hand-outs

Language teachers who have received training in corpus linguistics can resort to DDL as a supplement to their conventional teaching in two ways – direct or indirect. This means that learners can use concordances indirectly through corpus-based materials designed by teachers as handouts or they can have direct computer-based experience with corpora. The direct and indirect approaches are also termed as 'hard' and 'soft' [Gabrielatos, 1-37], or 'hands-on' and 'hands-off' [Boulton, *On DDL and Language Pedagogy,* 15-36] approaches, respectively.

According to Johns [*On DDL Examples,* 1-16], hands-off corpus driven activities can be introduced at lower levels of language proficiency for immediate results. They require minimal or no corpus training, which can be an advantage for learners who are reluctant to work with software or are not well aware of how to work with it or how to interpret the results [Boulton, *On DDL and Language Pedagogy,* 15-36]. However, the mere fact that learners work with a corpus-based handout does not guarantee successful learning unless the teacher is able to use them judiciously [Frankenberg-Garcia, 128-146]. For example, if the teacher's randomly selected concordance lines ask learners to infer the meaning of a random word from context, learners might find it frustrating assuming that they could more effectively find the meaning of the word in a dictionary. Instead, the concordances can be used to reinforce the meaning of the word or expand learners' previous one-off contact with the word. On the one hand, corpus based handouts can help learners avoid scrolling down countless concordance lines, when they have to read unedited texts and cannot decide what to look for. On the other hand, they will not be able to develop competency in using corpora [Frankenberg-Garcia, 128-146]. The soft type of DDL can be a solution in contexts where computers are not available at regular basis, valuable classroom time can be wasted because of the lack of technical back-up or inappropriate searches, and both teachers and learners are overwhelmed by the use of "new material (the corpora), new technology (the software), and new approach (DDL) all at once" [Boulton, *On DDL and Language Pedagogy,* 15-36]. Despite the overstated motivating factor of technology in education, computers can be unappealing for many teachers, as well as learners, and, therefore, become an obstacle for wider uptake of DDL.

The use of prepared materials allows the teacher to tailor activities to learners' needs and abilities [Boulton, *On DDL and Language Pedagogy,* 15-36] and avoid the indiscriminate use of concordances [Frankenberg-Garcia, 128-146]. This way the teacher can edit the language by leaving out the difficult language, by excluding offensive or sensitive language, etc., thus sheltering learners from many problems of working with raw corpus data [Frankenberg-Garcia, 128-146). Furthermore, printed materials can provide a gentle lead-in to hands-on experience [Gabrielatos, 1-37], and "scaffolding can be gradually reduced until students can be presented with concordance output to investigate independently and unaided" [Johns, *On DDL Challenge,* 107-117].

The difference between direct and indirect DDL is not merely the medium of delivery, but more than that. Hands-off concordancing does not have the full potential of hands-on corpus work. The latter can be achieved through extensive training, though it is often difficult to implement in an already established syllabus. The benefits that learners can extract from hands-on corpus consultation include flexibility, autonomy, lifelong learning, and long-term recall [Boulton, *On DDL and Language Pedagogy,* 15-36]. Direct corpus use can also provide learners with an experience of a linguist. However, there is a doubt as to whether it is necessary and a suggestion is given that hands-on corpus activities, like handouts, should be immediately applicable to learners' language learning interests, needs, and goals [Frankenberg-Garcia, 2014]. Through hands-on experience, learners have more opportunity to find answers to their individual questions, to select data relevant to them, to see more contexts, which are selectively printed on handouts.

The survey of 80 evaluations of DDL studies, conducted by Boulton [*On Learning Outcomes,* 129-144], revealed that most researchers favor computer-based corpus work, while only four studies focused on paper-based work. While studies report various findings on the learners' gains from DDL, Boulton suggests that they should be treated with caution,

meaning that several factors need to be considered – practical, cultural, individual, and pedagogical. Both hands-on and hands-off DDL have benefits and limitations, hence each might be appropriate for certain learners, teachers, and contexts. He also proposes that meta-analysis would provide data that could extend the confines of acceptability of the research outcomes.

The availability and use of a computer and a corpus in the language classroom is not enough for DDL [Reppen, 13-21]. The "corpus-informed language pedagogy" [Braun et al., 5], which comprises all the complexities of the field, requires three important steps, which allow both teachers and learners to avoid the pitfalls of DDL and successfully implement it – (i) careful selection of a corpus, (ii) awareness of corpora design, and (iii) skills and knowledge of its correct use. First, the choice of a corpus needs to be made with consideration of a number of factors, including learners' age, educational background, time period, genre of texts, etc. Second, the teacher needs to raise awareness of how a corpus is designed, which is essential for preparing both hands-on and hands-off activities. Corpora exploration is carried out through a concordancing program, which is typically used to conduct searches for a word or a group of words in different formats – as a frequency list, key word in context (KWIC), collocations, part of speech tagging (PoS), and so on [O'Keeffe, Farr, 506-517]. Third, the exploitation of a corpus demands certain skills and knowledge on the part of the teacher, which enable him/her to conduct corpus-based explorations, to receive more language-related insights and to evaluate corpus results in light of the preset pedagogical goal. In this respect, O'Keeffe and Farr [412] conclude that "The more teachers know about corpora and how to use them, the more they will be empowered to evaluate corpus-based materials objectively". Regarding learners' role, it is no less important for them to understand the benefits and know-how of corpus use, which will lead to more engaging and cognitively conscious language learning process.

The potential theoretical advantages that DDL promises are summarized by Boulton and Cobb [*On Meta-Analysis,* 348-393] in the following way:

1) DDL reflects current language theory. The latter views language as dynamic, interactive, complex, and patterned, as opposed to the view of rule-governed language [Tomasello, 408]. In light of this view, Taylor [384] describes language knowledge as a mental corpus of combined experiences of language use. Corpus linguistics provides insights into this patterning, including lexical priming, norms and expectations, and the idiom principle. With the help of DDL, learners are exposed to this patterned use of authentic language in context [Boulton, *On DDL and Language Pedagogy,* 15-36].

2) DDL reflects current learning theory. Rules are artificial mental abstractions, hence hard to acquire, in contrast to patterns, which our brain is programmed to notice around us [Barrett, Dunbar, Lycett 448]. Constructivist learning theory is in line with this approach and facilitates the acquisition of the target norm through progressive approximations. The role of DDL, in this respect, is significant as it fosters autonomous and lifelong learning skills that are transferrable to new contexts and enhance learning [Boulton, *On DDL and Language Pedagogy,* 15-36].

3) DDL reflects current psycholinguistic theory. Being a natural process, pattern induction minimizes the load of cognitive processing [Sweller, 37-76] on the one hand, and still requires cognitive effort for constructing meaning, on the other hand. This effort, which is a reliable factor for retention, is absent in rule-based instruction and is required by DDL, when learners are exposed to multiple patterned examples made salient in authentic input necessary for noticing [Schmidt, 1-63].

4) DDL reflects current second language acquisition research. It provides mediation along the continuum of meaning-focused and form-focused language instruction, as well as top-down and bottom-up processing – recommendation that has been theorized for years but not practiced by language educators.

5) DDL reflects current learner practice. The world wide web can be considered a huge "corpus" that allows learners to find answers to their questions by googling as "concordancing". Thus, DDL activities that reflect this practice can refine it and

progress to corpus work [Boulton, *On DDL and Language Pedagogy,* 15-36].

It is only through direct contact with corpora that learners gain such long-term benefits as language awareness, noticing and autonomy [Boulton, *On Meta-Analysis,* 348-393].

Conclusion

The paradigm shift in language learning and teaching, brought by a number of researchers resulted in DDL, which brings together theories of constructivism, communicative approach, and advances in the field of autonomy. It is based on the key concepts of authentic data, learner-control, discovery learning, autonomy, and revolutionaries. Many corpus-based studies have been carried out in language learning, but language corpora have not been integrated into mainstream teaching practices. Researchers point to the need to shift the use of DDL from a research-oriented process to a more pedagogically underpinned one. As discussed above, a lot still needs to be done before corpora can actually be implemented in language pedagogy. One reason that language corpora have not become part of mainstream language instruction is the lack of teacher training and resources. Another reason could be the lack of attention paid to learners' attitudes towards using language corpora, particularly at low levels of language proficiency, and to the language areas that would benefit most from DDL. This would help to mitigate the risks, fears, and challenges of using DDL and anticipate more benefits in language pedagogy.

References

1. Barrett, L., Dunbar, R. and Lycett, J. Human revolutionary psychology. Basingstoke, UK: Palgrave, 2002, p. 448.
2. Kennedy, G. An introduction to corpus linguistics, London: Longman, 1998, p. 328.
3. Mishan, F. Designing authenticity into language materials, Bristol and Portland: Intellect Books, 2004, p. 346
4. O'Keeffe, A., McCarthy, M. J., and Carter, R. A. From corpus to classroom: language use and language teaching, Cambridge: Cambridge University Press, 2007, p. 314.
5. Taylor, J. The mental corpus: How language is represented in the mind, Oxford, UK: Oxford University Press, 2012, p. 384.
6. Tomasello, M. Constructing a language: A usage-based theory of language acquisition, Cambridge, MA: Harvard University Press, 2003, p. 408.
7. Bernardini, S. "*Exploring new directions for discovery learning"*, in Kettemann, B. and Marko, G., eds., Teaching and learning by doing corpus analysis, Proceedings from the Fourth International Conference on Teaching and Language Corpora, Graz 19-24, July 2000. Amsterdam: Rodopi, 2002, pp. 165-182.
8. Boulton. A. "*Data-driven learning and language pedagogy"*, in Thorne, S. and May, S. eds., Language, Education and Technology: Encyclopedia of Language and Education, New York: Springer, 2017, pp. 15-36.
9. Boulton, A. and Cobb, T. "*Corpus use in language learning: A Meta-Analysis"*, Language Learning, 67(2), 2017, pp. 348-393.
10. Boulton, A. "*Learning outcomes from corpus consultation",* in Moreno, J., M., Valverde, F. S. and Pirez, M. C., eds., Exploring new paths in language pedagogy: lexis and corpus-based language teaching, Londres: Equinox, 2010, pp. 129-144.
11. Braun, S., Kohn, K. and Mukherjee, J. "*Corpus technology and language pedagogy: New resources, new tools, new methods"*, English Corpus Linguistics, Volume 3, 2006, p. 214.
12. Chambers, A. and Kelly, V. "*Semi-specialized corpora of written French as a resource in language teaching and learning"*, Teanga, 21, 2002, pp. 20-21.

13. Flowerdew, L. "*Data-driven learning and language learning theories: Whither the twain shall meet",* in Leńko-Szymańska, A.and Boulton, A., eds., Multiple affordances of language corpora for data-driven learning, Amsterdam: John Benjamins, 2015, pp.15–36.
14. Frankenberg-Garcia, A. "*The use of corpus examples for language comprehension and production"*, ReCALL, 26(2), 2014, pp. 128-146.
15. Gabrielatos, C. "*Corpora and language teaching: Just a fling, or wedding bells?"*, TESL-EJ, 8(4), A-1, 2005, pp. 1–37.
16. Gaskell, D. and Cobb, T. "*Can learners use concordance feedback for writing errors?"* System, 32(3), 2004, pp. 301–319.
17. Johns, T. "*Should you be persuaded: two examples of data-driven learning"*, English Language Research Journal, 4(1), 1991, pp. 1–16.
18. Johns, T. "*Data-driven learning: The perpetual challenge"*, in Kettemann, B. and Marko, G., eds., Teaching and learning by doing corpus analysis. Amsterdam: Rodopi, 2002, pp. 107–117.
19. Johns, T. "*Micro-Concord: a language learner's research tool"*, System, 14(2), 1986, pp. 151-162.
20. O'Keeffe, A. and Farr, F. "*Using language corpora in language teacher education: Pedagogic, linguistic and cultural insights"*, TESOL Quarterly, 37(3), 2003, pp. 389–418.
21. O'Sullivan, I. "*Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy"*, ReCALL, 19(3), 2007, pp. 269–286.
22. Reppen, R. "*Building a corpus: what are key considerations?"*, in O'Keeffe, A. and McCarthy, M. J., eds., The Routledge Handbook of Corpus Linguistics, 2nd Ed. London: Routledge, 2022, pp. 13-21.
23. Schmidt, R. "*Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning"*, in R. Schmidt, ed., Attention and awareness in foreign language learning, Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center, 1995, pp. 1-63.
24. Sweller, J. "*Cognitive load theory"*, in J. P. Mestre & B. H. Ross, eds., The psychology of learning and motivation: Cognition in education, 2011, pp. 37–76.
25. Widdowson, H.G. "*On the limitations of applied linguistics"*, Applied Linguistics, 21, 2002, pp. 2-25.

Ներկայացվել է՝ 16.10.2022թ.
Ուղարկվել է գրախոսման՝ 04.11.2022թ.