# 1 What Is Data-Driven Learning?

## 1.1 Introduction

This Element has been written for language teachers, graduate students, and professionals in Teaching English (or other languages) to Speakers of Other Languages (TESOL), language education and applied linguistics with little or no prior experience in corpus linguistics. We have aimed to keep things simple and accessible, introducing the basics step by step. This Element discusses the use of corpora and other tools to enhance language teaching and learning, highlighting the importance of authentic texts and inductive learning – learning by exploring the language with appropriate tools, as opposed to 'being taught'. The Element has been designed as a practical guide for those wishing to find out more about Data-Driven Learning (DDL). It provides a general overview of the use of corpora in language education as well as more specific discussions about ways to implement DDL across different teaching scenarios.

DDL is an approach to language teaching and learning based on the tools and techniques of corpus linguistics to enhance learners' understanding of how language is used across linguistic genres. A corpus is usually an electronic collection of authentic texts designed to be representative of a language or language variety (e.g. Spanish, or contemporary teenage speech, or research articles in a given scientific discipline). The point of DDL is for foreign or second language (L2) learners to explore how language actually works, as opposed to the traditional 'being taught'. DDL favours inductive learning and general cognitive skills such as critical thinking, pattern detection, hypothesis building, data analysis, and noticing. Thanks to the large and increasing body of research in this area, we know that DDL promotes learner autonomy and language awareness.

A corpus (plural *corpora*) often contains millions or billions words, written or spoken by hundreds or thousands of different people, thus providing an insight into real usage over a wide range rather than the intuitions of a few individuals who may produce coursebooks or other materials. For example, the British National Corpus (BNC)[1] was designed in the late 1980s to represent the English language used in the UK. More recently, the Corpus of Contemporary American English (COCA)[2] was built to illustrate English as used in the United States of America. The former contains 100 million words, the latter 1 billion in 2024, and is still growing. However, not all corpora are this big. Smaller corpora can be useful to study very specific varieties of language use (e.g. research papers in the field of dentistry), or to provide an example of an actual piece of

[1] www.english-corpora.org/bnc.
[2] www.english-corpora.org/coca.

research (e.g. the articles published about a celebrity during a period of time in a given newspaper).

Human beings create and shape their understanding of reality through the use of language, which involves engaging in different forms of communication. This can include spoken interactions, such as conversations and speeches; written formats, like books, articles, and social media posts; and hybrid modes, which combine elements of both, such as digital communication that may involve text, audio, or visual content simultaneously. Through these discursive practices, individuals interpret their experiences, share their thoughts, and construct the meanings that define their social and personal realities. Corpora are resources that facilitate the study of how the speakers of a language use it across different linguistic genres such as fiction, news, conversation, TV shows, and academic language. Corpora, in essence, provide evidence of how human beings construct their reality discursively by engaging in spoken, written or hybrid communication modes. Corpora, therefore, have the potential to show how a community of speakers uses language to fulfil real communicative functions, presenting contextualised authentic language usage. Corpora have been used to build dictionaries based on real use of the language, to study the grammar of English, to understand how humans generate discourses based on ideologies or, just to name one other area of application, to attribute authorship in forensic linguistics.

A list of online resources and software for DDL specifically compiled for this Element can be found at https://doi.org/10.48316/DgLhT-C33sa (Pérez-Paredes & Boulton, 2024).

One of the most important uses of corpora is in language education. Before the widespread availability of personal computers, a few teachers in the late 1960s were already making use of printed concordance lines (McEnery & Wilson, 1997) to explore actual language patterns in English for Specific Purposes (ESP). It was this possibility that sparked the interest of these pioneering teachers in what concordance lines could bring to language learning. This is useful to both language teachers and learners, who cannot always find frequency-based evidence of usage in dictionaries or textbooks. We will find some examples of these uses in Sections 3 and 4, where we will discuss the use of corpora in secondary-school contexts and in academic writing teaching contexts, respectively.

In this Element, we will examine language learners' use of and engagement with concordance lines such as those in Figure 1. Each line is a short extract of text from a corpus that shows a word or phrase in its immediate context, and can easily be printed out to create activities. For reasons of space, Figure 1 shows just a few of the hundreds of lines available in the corpus. The focus word or phrase (here *sort of*) remains in the centre of the line, allowing the user to read the words that appear


**Figure 1** *sort of* concordance lines from the 2014 Spoken British National Corpus

both to the left and to the right and see how the word or phrase is used in context. These concordance lines are a random selection taken from a pool of 14,437 occurrences of *sort of* in 10.5 million words of spoken English in the 2014 British National Corpus (BNC).[3] Each of those occurrences is an attested, authentic use that can be situated in a specific text and context. Exploring these lines involves detailed, usage-driven observation of the contexts where the phrase appears in spoken language, and the words that usually accompany their use (e.g. *thing*, *like*, *stuff*, *know*).[4] The concordance lines also show that *sort of* in spoken British English is followed by a singular noun in almost 30 per cent of the instances and by a plural noun in only 6 per cent. Adjectives are found after *sort of* in 6 per cent of the attested uses; *–ing* verbs and verbs in the simple past are more frequent than any other verbal forms. Although nowadays concordance lines are always generated by downloadable software or by online platforms (see Section 3), using printed concordance lines in the classroom has been, and continues to be, a standard way to show learners how words are combined in actual usage.

This type of detailed analysis of 'words in context' (lexical items embedded in strings of natural discourse) highlights the frequent patterns of how language is typically used: in other words, we are not dealing with grammar rules that show what is possible, but normal use that shows what is probable. It more closely reflects our current thinking of how language is learned, stored and used (usage-based theories; see e.g. Tomasello, 2003), and the DDL approach has been shown to increase language learners' awareness about the structure of the language, and language teachers' and material developers' sensitivity to the need to incorporate such description of usage in language teaching materials

[3] https://cass.lancs.ac.uk/bnc2014.

[4] In linguistics, 'usage' refers to how language is actually used by speakers and writers in real-world contexts, as opposed to how it might be prescribed by formal grammar rules or traditional standards. Studying usage involves observing and describing the patterns, choices, and variations people make when communicating. This includes vocabulary selection, grammar structures, pronunciation, and even stylistic choices, reflecting the dynamic nature of language.

(see Adolphs & Carter, 2003, for an analysis of *like* in conversation and a proposal to incorporate those findings to the *Cambridge Advanced Grammar of English*).

## 1.2 Data-Driven Learning and the TaLC Community

In DDL, discovering language patterns becomes the quintessential activity for both teachers and learners. Tim Johns coined the term *Data-Driven Learning* in 1990 to describe the type of corpus-driven work that the author himself encouraged in a group of international students in the UK when learning grammar. Some of the questions that came up during his sessions included 'What is the difference between *therefore* and *hence*?' or 'Why aren't all *shoulds* real *shoulds*?' In Johns' words, concordance lines allow the teacher to abandon the role of expert and take on that of a research organiser to help in the discovery of language patterns, embracing the 'I'm not sure: let's find out together' maxim (Johns, 1990, p. 31). To answer this type of question, it was necessary to examine the concordance lines generated by a specific type of software: a concordancer. Four years earlier, Johns (1986) had created *Micro-Concord*, software that allowed language teachers and learners to generate concordance lines on a Spectrum computer. It is fascinating to think that almost 40 years later, we are living the early years of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs). We will come back to this point in Section 5.

In 1994, the first edition of the Teaching and Language Corpora (TaLC) biennial conference series was held at the Lancaster University in the UK. Since then, it has brought together researchers, lecturers, language teachers and other professionals interested in the use of corpora in language education. TaLC has served as a catalyst for the discussion of teaching practices in DDL, promoting the publication of volumes that feature cutting-edge research and workshops that showcase the latest developments in software, applications and classroom practices. The latest TaLC-inspired volumes (Charles & Frankenberg-Garcia, 2021; Götz & Mukherjee, 2019; Leńko-Szymańska & Boulton, 2015; Pérez-Paredes & Mark, 2021; Tyne & Spina, in press) include discussions about, among other things, language learners' engagement with corpora and corpus tools, the design of DDL activities, and how analyses of learner language can inform teaching practice. TaLC is one of the largest, if not the largest, community of researchers and teachers worldwide that discusses uses of corpora for the teaching of languages across a variety of learning contexts and a great meeting point to discuss the affordances and the challenges behind the use of corpora for language teaching.

There have been thousands of academic publications on DDL over the last 40-plus years, of which at least 800 provide some kind of empirical evaluation.

Because it can be difficult to make sense of such large quantities of research, various researchers have produced a number of syntheses. The broadest to date is Boulton and Vyatkina (2021), which coded 489 papers to examine how the main areas covered have changed over time. They followed this up with a more focused look at DDL for English language teaching, based on papers in highly cited journals (Boulton & Vyatkina, 2024). Pérez-Paredes (2022) explored key clusters from five journals over a five-year period, arguing that DDL needs to become 'normalised' – that is, part of the everyday practice in language classrooms. Dong et al. (2023) used a bibliometric analysis to identify the principal research themes and contributors to the DDL paradigm. Most relevant for present purposes are probably the various meta-analyses that have been carried out, a meta-analysis being a method for combining quantitative results from different studies. Though meta-analyses are not without their drawbacks, they do afford a broader view as they enable the pooling of many studies into a single whole, with each of the included papers featuring data from different types of DDL in different contexts, with different learners, corpora, tools, and procedures. Boulton and Cobb (2017) included sixty-four studies in the field DDL as a whole, finding large effect sizes overall (i.e. DDL is in the top quartile, providing better results than 75 per cent of other meta-analyses in second language acquisition according to Plonsky and Oswald (2014))". Like most meta-analyses, they also divided the sample into groups to explore different 'moderator variables' (e.g. comparing hands-on concordancing and paper-based DDL, or with learners at different levels of proficiency), concluding that 'DDL works pretty well in almost any context' where we have sufficient data (p. 386). Ueno and Takeuchi (2023) followed similar procedures for more recent research, finding medium effect sizes but their methodology is not beyond criticism (see Boulton et al., in press). Lee et al. (2019) focused on DDL for vocabulary, where the twenty-nine studies included showed large effect sizes for learning both referential meanings and syntactic features. Most recently, Ngo and Chen (2024) looked at thirty studies of corpus use in ESL/EFL writing, again with large effect sizes. The conclusion: DDL works, and it works well! But like all studies in language teaching and learning, it varies in how well it works according to any number of factors – the learners' age, proficiency, needs and aims, the tools and corpora used and the activities to accompany them, and so on. DDL has potential; whether it will work for you and your students is something you can only try.

## 1.3 Linguistic Data and Language Teaching

DDL seeks to foster learners' inductive learning where authentic language data from corpora provide opportunities for discovery and internalisation of usage in an effective way (Boulton & Vyatkina, 2021). This is well reflected in

collections of DDL activities such as Viana (2022) or Le Foll (2021), which have brought together classroom resources that aim to provide teachers with ready-to-use activities for exploring formulaic language, collocations and language patterns, among other things. In the following sections, we will examine DDL activities as well as corpora and strategies to use corpora in the classroom.

Researchers in language learning have shown that both explicit and implicit instruction methods have a significant impact on L2 learning, although it seems that implicit instruction shows longer-lasting effects on learning compared to explicit instruction (Kang et al., 2019). While in explicit L2 instruction learners are taught grammar or vocabulary through explanations or drills, implicit language learning involves exposure to language input without overt instruction on grammar rules or language structures. According to the meta-analysis on form-focused instruction carried out by Kang et al. (2019), implicit learning takes place through pattern recognition, incidental learning, repeated exposure to language input and unconscious cognitive processes. DDL is well positioned to favour a wide spectrum of implicit and explicit language learning (O'Keeffe, 2021), encouraging attention to linguistic form. In DDL, language learners examine linguistic evidence in terms of words, word combinations and patterns (Hunston, 2019), and formulate their own hypotheses about how language works, which may lead to the development of language skills and knowledge. This constructivist approach means that learners are deriving their own 'rules' (i.e. formulations of language patterns) from the language they encounter. While their rules may be less accurate than those a teacher could provide (though this is not always the case!), the process requires significant thinking (cognitive depth), and the rules will be meaningful to each individual; for both these reasons, they are more likely to be retained.

In DDL, teachers and learners interact with two types of language data. The first type of data is the corpus itself. A corpus can be conceptualised as a group of texts that have been collected to increase our understanding of how language is used. Understanding the composition of the corpus helps to make sense of the data that can be extracted from the corpus. As McEnery and Brezina (2022) have put it, knowing the purpose of the corpus, the language varieties included, and any biases in the data collection process helps users interpret the linguistic phenomena observed in the analysis. A well-designed corpus can reveal the discursive practices of the speakers of a language, either because the corpus represents the main genres they use (e.g. fiction, blogs, and conversation), or because the corpus focuses on one specific genre across time.

The second type of linguistic data in DDL is the insights gained by examining a corpus. These insights fall within one of the following four categories: frequency measures, collocations, language patterning, and genre awareness

(Pérez-Paredes, 2022). Table 1 offers a summary of what these insights mean for language teachers and learners.

More broadly, Boulton and Cobb (2017, pp. 350–351) argue that DDL aligns with current theories in a number of fields:

- Linguistics: rather than being rule-based, language is probabilistic, dynamic and complex, interactive and patterned; knowledge is based on an amalgam of all previous encounters.
- Learning: rules are hard because they are an 'artificial intellectual abstraction', while patterns are easier, reflecting an innate ability to make sense of the complex world around us.
- Psycholinguistics: while pattern detection involves deep cognitive processing, the naturalness of it reduces cognitive load, thus freeing up space for attention to meaning.

**Table 1** Insights from corpus data consultation and analysis

| | |
|---|---|
| Frequency measures | Frequency information can help language learners and teachers prioritise which words and phrases to focus on for vocabulary acquisition. High-frequency items are more likely to be encountered in real-world communication and are essential for building a solid language foundation (Szudarski, 2022). |
| Collocations | Words occur together in natural language in non-random ways. Learning and using collocations can make learners' L2 more natural and idiomatic, contributing to fluency (Friginal & Roberts, 2022). |
| Language patterning | Awareness of language patterns enables language learners to communicate more effectively and fluently. If learners understand and use language patterns to convey meaning accurately, they are more likely to produce language that sounds natural (Gablasova & Bottini, 2022). |
| Genre awareness | Corpus data and frequency analysis allows teachers and learners to identify variations in language use across genres (e.g. business emails or lab reports) or language varieties (e.g. Spanish in Mexico or Spain). Frequency information across genres provides insights into what is preferred by the users of different genres and facilitates our understanding of how vocabulary and patterning work in a given genre. |

- Second language acquisition: effective learning requires a balance between top-down processing (meaning in discourse and context) and bottom-up (from sounds, words or grammar to create meaning), with language at the centre of communication.

DDL also reflects existing learner practices, involving aspects of computers that learners are already familiar with and doing in their own time (searching for answers to their questions and interpreting the results they find on the internet). In fact, various authors have attributed all kinds of attributes to DDL, such as authenticity, autonomy, communication, consciousness-raising, constructivism, context, critical thinking, discovery learning, dynamic systems theory, focus on form, heuristics, Information and Communication Technology (ICT), individualisation, induction, input flood, languaging, learner-centeredness, learning-to-learn, life-long learning, Languages for Specific Purposes (LSP), (meta-)cognition, meaningfulness, mobile learning, motivation, needs, noticing, responsibility, salience, scaffolding, sensitisation, strategy training, styles and preferences, tasks, transferability, and so on. We will not develop all these here: suffice to say that there are numerous reasons to think that DDL has much to contribute to L2 learning and use.

## 1.4 DDL in the 2020s and Beyond

DDL practices have undergone considerable changes in the last two decades. These changes have affected the types of resources for DDL as well as the scope of classroom practices. One of the most visible is a shift from small corpora of up to 1 million words to corpora of billions of words. There have also been shifts from desktop-software-only applications to web corpus management tools that facilitate access anywhere, any time; from workstations and labs for hands-on learning to students' use of their own personal devices, including tablets. Additionally, there has been a continuous interest in expanding DDL to other languages and other curricula, including secondary education and professional situations. In essence, DDL has become more open to a wider range of teaching contexts, languages, and tools.

One of the areas where DDL has undergone change is the use of a variety of resources in language classrooms, not just corpora. As we will discuss in Section 3, DDL has made use of texts and collections of texts that were not meant to represent language varieties or to support linguistic research. Such collections tend to be smaller and to show some flexibility in terms of their composition. Classroom resources do not necessarily make use of corpora that are representative of a language or variety, but of the type of language which is needed by language learners. Such observations prompted the idea of Broad

Data-Driven Language (BDDL). BDDL (Pérez-Paredes, 2024) can be implemented more straightforwardly in classrooms as the corpus resources are either typically put together by teachers, which makes them in principle more readily accessible to learners or are simply available on the internet. Such resources include, among many others, collocation dictionaries, collocation tools, phrase extraction tools, grammar pattern identification tools, and n-gram generators.[5] This list highlights some important concepts in corpus linguistics. First, collocates are words that occur together more often than you would expect by chance. One example is *blond* which typically collocates with very few items in English – *hair* (including *curls*, *beard*, etc.) or *people* with such hair (*woman*, *child*, etc.). This may seem obvious, but such patterns of usage are not identical between languages: customers frequently ask for *une bière blonde* in French, where English would prefer another term such as *lager* over *blond beer*. The concept of *phrase* needs little explanation, but corpus linguistics has no notion of meaningful groups. What it can do however is to detect *n-grams*, which is a sequence of *n* items (usually words), regardless of whether they constitute a meaningful unit (See Section 2). So *one of* is a 2-gram (or bigram), *one of the* is a 3-gram, *one of the most* is a 4-gram, *one of the most important* is a 5-gram, *one of the most important things* is a 6-gram, and so on.[6] Different people use different labels for these groups (e.g., clusters or chunks), but the point is that they show the patterns in language and how they can be built up, broken down, and where variation is more or less likely.

BDDL enhances language learners' engagement with language, facilitating their analysis of linguistic patterns and structures, and identification of recurring word combinations or collocations. The interpretation of language data is essential for extracting meaningful information from corpora and other language-driven sources, and for applying it to language learning. This is, in our view, a core feature of data/corpus literacy that will be essential in new ecologies of digital language learning that have already emerged (Gee & Hayes, 2011) and which are likely to be shaken by the impact of LLMs on language learning. In Section 3, we will discuss in detail how corpora and GenAI afford data analysis and language learning.

In a context where natural language processing technologies and GenAI evolve rapidly and where we can expect massive changes in the way we interact with computers, interpreting language data for language learning is a core component of new digital literacies via GenAI. Working with corpora and

[5] A list of resources for English can be found on www.perezparedes.es/selected-corpus-resources-for-second-language-educators.

[6] *One of* is among the most frequent 2 grams in COCA, which is most frequently followed by *the* and then the other words in each of the examples given.

using DDL activities can help our students reframe how language data is generated and presented, and how it can be interpreted in language education. In this context, learners will have to:

- Understand data concepts. These include types of data, structured (e.g. a table with headings and/or variable names) vs. unstructured datasets (e.g. running text), data sources, and data formats.
- Interpret data. This involves frequency and its impact on language use, including basic understanding of charts and tables and drawing conclusions from data visualisations.
- Analyse data. This draws on basic concepts from corpus linguistics (cf. the 20 basic Corpus Linguistics skills in Pérez-Paredes, 2020).
- Communicate data. Talking about language data can help students process frequency-based information and can help teachers scaffold activities that can enhance their understanding of how learners can infer language properties from the data presented in corpus consultation or other soft DDL tools discussed above.

## 1.5 This Element

The initial challenges of DDL for beginners, such as preparing corpora and using tools, can feel intimidating. Addressing these obstacles could further ease practitioners into DDL. Our overall aim is to equip language teachers and graduate students in TESOL, applied linguistics and language education programs with skills to promote language awareness and critical thinking in their students. The Element has been structured to facilitate the progressive acquisition of knowledge about DDL and the use of corpora in language teaching. Sections 2 to 5 each emphasise distinct themes, but many of these can be useful in other areas; for example, the discussion about using DDL with young learners in Section 3 leads to an examination of the different roles that language teachers may adopt when using corpora; similarly, the discussion of spoken data or learner corpora in Section 5 is potentially relevant in other situations. The reader is encouraged to tailor their reading experience according to individual preferences. The structure of the Element features six sections including this one. The next section covers the essentials, while Sections 3, 4, and 5 each introduce specific scenarios:

- **Section 2** surveys various existing corpora and their applications in DDL. It outlines different DDL activities and the use of concordancers – tools that help analyse language patterns found in the texts. The section also discusses the basics of querying corpora in the classroom, providing