CHAPTER 2

Data-driven learning, theories of learning and second language acquisition

In search of intersections

Anne O'Keeffe
Mary Immaculate College, University of Limerick

This chapter focuses on the need to address both theories of learning and theories of language acquisition in data-driven learning (DDL) research. While it recognises that there has been so much worthwhile research work on DDL which has shed so much light on the value of DDL, it is still not a mainstream methodology. The chapter argues that by understanding better the variations in pedagogical underpinnings and ontologies, DDL research can better pinpoint what works within specified variables. Additionally, the paper argues strongly for engagement with ongoing research in second language acquisition (SLA), especially from a usage-based perspective because there are so many resonances for DDL in terms of the centrality of the role of frequently experienced syntactic regularities in learning.

Keywords: data-driven learning, second language acquisition, usage-based theory

Introduction

In 2006, Mukherjee cogently summarised the state-of-the-art in relation to corpus linguistics and language pedagogy, noting that in spite of the "undeniably large number of corpus-based activities that have been suggested by researchers in applied corpus linguistics", there is a gap "between what applied corpus linguistics has to offer and what teachers actually do (or don't do) with corpora in their teaching" (2006, p. 20). In addition to enhanced teacher professional development, he pointed to the need for corpus activities to be evaluated under real-time conditions in real classroom contexts and both the perspective of the teacher and the learner. It is interesting to look back on this paper to see a point in time where there was a need to make such a call in relation to data-driven learning (DDL). It indexes a time when the number of classroom teachers who used corpora were few and far

https://doi.org/10.1075/scl.102.02oke

between and Mukherjee called for teachers "to be involved to a much larger extent in corpus-based classroom action research" (2006, p. 20).

Mukherjee's (2006) call, it is fair to say, was of its time. More than a decade on, we can say that we know so much more about using corpora in real classrooms as a result of so many real teachers in real classrooms exploring DDL in their practice (albeit mostly in the context English in higher education settings). Recent meta-analyses of research on DDL not only offer a summative overview of the many insights that have accrued from such studies, they also point to methodological weaknesses to be addressed within this seam of research (see for example Boulton & Cobb, 2017; Lee et al., 2019; Pérez-Paredes, 2019). Reviews and meta-analyses draw together the key outcome variables that have been measured across the gamut of DDL research (within set criteria), over a period time. In doing so, they also tell us about the variables that have not been a key concern in DDL studies.

This paper argues that while we have gained so much insight from existing work on DDL, we have largely not engaged with theoretical concerns and this may be to the detriment of embedding DDL as a more mainstream pedagogy. It is argued that there is a need for (1) greater refinement of the pedagogical side of DDL research, and (2) greater connection with relevant theories of second language acquisition (SLA), especially emerging work on the connections between SLA and corpus data. Undoubtedly, this needs to be a two-way process where the findings of DDL also enrich instructed SLA models and perspectives.

It is important to note that this paper does not set out to undermine the existing work on DDL. Its goal is to open up new debates and motifs within our research community. It is hoped that the call inherent in this paper will influence the DDL research narrative and lead us to an enhanced rationale for the benefits of using DDL in the classroom, one which is based on and interlinked with theories of learning and second language acquisition.

It is acknowledged that this paper is not making a new call. In the past, authors such as Römer (2006); Tribble (2008) and Pérez-Paredes (2010, 2019), among many others, have pointed to the need to find a plausible way of moving DDL from a research-oriented process suited to university settings to one with a broader pedagogical application and underpinning. More recently, O'Keeffe (2020) makes a similar call for DDL research to broaden its research gaze so as to address underlying theories of both pedagogy and language acquisition. As Römer (2006, p. 129) puts it, much still needs to be done before we can say that "corpora have actually arrived in language pedagogy". In addition to Mukherjee (2006), this paper is very much influenced by work from over a decade ago, by eminent corpus linguist Stig Johansson who wondered why the potential of DDL for enhancing language learning was not being realized (Johansson, 2009), especially given the parallels that he could see with SLA research that was ongoing at that time. Johansson (2009)

saw connecting with SLA concepts such as input enhancement and the role and nature of attention as some of the obvious nexuses for our research community. Johansson (2009) makes the case that DDL is well-placed to conduct research that could lead to cutting-edge insights that can enhance ongoing SLA debates, especially in relation to implicit and explicit learning processes. Another major influence on this chapter is Flowerdew's thought-provoking (2015) paper which notes, over the years, that DDL and SLA research paths have run in parallel and have rarely intersected.

Theoretical positions and motifs in DDL

In terms of theories of learning, as noted in O'Keeffe (2020), there are two main motifs in DDL research: (1) those who take a constructivist perspective, and (2) those who express or allude to a more socio-cultural model of learning. The former places its emphasis on discovery learning while the latter lauds the benefits of mediated learning (through opportunities that teacher-mediation or peer-to-peer learning offer). The original spirit of DDL was certainly a constructivist one where learners engage in and gain from cognitively grappling with data (Johns & King, 1991; Johns, 1994; Cobb, 2005; O'Sullivan, 2007; Boulton, 2010; O'Keeffe, 2020). As Mukherjee (2006, p. 11) notes, Widdowson's (1990) "learning as discovery" came through in the vanguard of DDL, where learner-centred inductive learning was fostered. Indeed, Boulton (2010) pointed out that constructivism fits well with DDL. The ideal of learners naturally consulting a corpus as a resource just as they would pick up a dictionary or a grammar reference book is a situation that we all aspire to for our language learners. As Boulton (2010, p. 535) succinctly puts it, this constructivist ideal means a situation where "learners are using adaptive behaviour in detecting regular patterns in the data that are meaningful to them, rather than attempting to learn and apply rules they are given, a more 'artificial' intellectual activity". This paper is not setting out to dispute this ideal but it is interested in unpacking it a little.

Going back to Johns and King (1991), a signature paper in the field, we find DDL 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 computer output" (Johns & King, 1991, p. iii). We need to ask ourselves: why is it important for learners to focus on 'regularities of patterning' so as to detect regular recurrences that are meaningful to them as an individual learner? Is this more beneficial than applying explicitly learnt rules of language? Smart (2014, p. 185) also points to the need to unpick our beliefs about the benefits of DDL when he notes that while there is evidence

of its benefits, it is not clear whether these accrue from the inductive approach to learning or to the use of corpus-informed tools and data (or a combination of all of these). Smart called for further research into *how* the inductive approach within DDL contributes to learning.

The other pedagogical motif, referred to above, that one finds in DDL literature relates to the degrees of mediation by the teacher (e.g. scaffolding) and the mediation between peers (e.g. working in pairs or groups). The degree to which a learner's interaction with the data (whether paper- or screen-based) is mediated by a teacher or a peer interrelates with degrees of autonomy and self-regulation. Such considerations bring us into the territory of sociocultural theory (SCT), where the role of mediation by the teacher, by peers and by the self in the process of learning are important variables (see Lantolf & Ahmed, 1989; Lantolf & Appel, 1994). SCT concepts feature much less overtly in the discourse of DDL but they certainly are there (see Kennedy & Miceli, 2001, 2017, for example).

The most common expression of SCT is through the concepts of self-regulation, teacher mediation and scaffolding, as well as peer-to-peer learning (see O'Keeffe, McCarthy & Carter, 2007; Huang, 2011; O'Keeffe, 2020). Indeed in the quotation above from Boulton (2010, p. 535), we see a reference to the SCT concept of self-regulation when he mentions "learners are using adaptive behaviour …" when consulting a corpus. The fact that DDL has different theoretical underpinnings, ranging from constructivist to sociocultural, is not in the least bit problematic, but it needs to be articulated and better understood by those using DDL and by those researching the use of DDL in the classroom. Most of all it needs to become one of the outcome variables of DDL research. In other words, as argued in O'Keeffe (2020), DDL studies need to work towards clarity of pedagogical position when undertaking research and interpreting results so that we can get a better insight into how pedagogical stance impacts on the classroom processes and the learning outcome(s).

Mukherjee (2006, p. 12) offered us a very useful framework for describing DDL activities. He proposed a cline based on learner autonomy, "ranging from teacher-led and relatively closed concordance-based exercises to entirely learner-centred corpus browsing projects". This notion of a cline is extremely useful. We can see the nature of the theoretical gap that we are dealing with if, for instance, at on one end of the cline, a pedagogical approach uses "serendipitous corpus browsing" (Bernardini, 2004, p. 22) while another uses controlled types of tasks such as illustrated below. In the following example (based on Poole, 2018, p. 13), the teacher directs the student as to the exact search to undertake. First, learners are directed to the following screen setting (Figure 2.1) and are asked to enter *beautiful* in one search box and *attractive* in the next. They are then asked to set the collocation window span to 0L to 1R (i.e. zero to the left and one to the right).

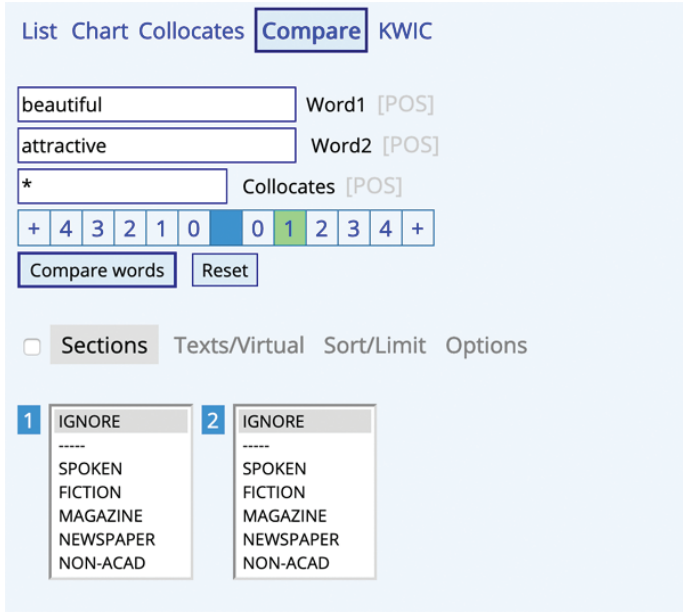

Figure 2.1 Example of a controlled corpus task based on Poole (2018, p. 13) using the *British National Corpus* within the BYU corpora interface

This setting within the BYU corpora interface generates the following results (at the time of writing). Students are directed to consider whether *beautiful* and *attractive* can be used interchangeably based on what they induce from the collocates list in Figure 2.2.

WORD 1 (W1): **BEAUTIFUL** (1.66)

	WORD	W1	W2	W1/W2	SCORE
1	DAY	45	0	90.0	54.2
2	HAIR	40	0	80.0	48.1
3	EYES	32	0	64.0	38.5
4	VOICE	27	0	54.0	32.5
5	BEACH	19	0	38.0	22.9
6	MORNING	17	0	34.0	20.5
7	EVENING	16	0	32.0	19.3
8	OBJECTS	16	0	32.0	19.3
9	SKIN	15	0	30.0	18.1
10	BRIDE	14	0	28.0	16.8
11	BAY	14	0	28.0	16.8
12	NAME	14	0	28.0	16.8
13	SUMMER	14	0	28.0	16.8
14	HEAD	12	0	24.0	14.4
15	FLOWER	12	0	24.0	14.4
16	SOUND	12	0	24.0	14.4
17	CHILDREN	11	0	22.0	13.2
18	SOUTH	20	1	20.0	12.0
19	CONDITION	10	0	20.0	12.0
20	BABY	10	0	20.0	12.0

WORD 2 (W2): **ATTRACTIVE** (0.60)

	WORD	W2	W1	W2/W1	SCORE
1	FEATURE	48	0	96.0	159.5
2	PROPOSITION	45	0	90.0	149.6
3	OPTION	29	0	58.0	96.4
4	ALTERNATIVE	25	0	50.0	83.1
5	FORCES	15	0	30.0	49.9
6	FOOTBALL	14	0	28.0	46.5
7	ENVIRONMENT	13	0	26.0	43.2
8	PRICES	10	0	20.0	33.2
9	TERMS	10	0	20.0	33.2
10	PRICE	9	0	18.0	29.9
11	PACKAGE	9	0	18.0	29.9
12	MEANS	9	0	18.0	29.9
13	INVESTMENT	7	0	14.0	23.3
14	PERSONALITY	7	0	14.0	23.3
15	PROSPECT	13	1	13.0	21.6
16	PROGRAMME	6	0	12.0	19.9
17	PRODUCT	6	0	12.0	19.9
18	OPTIONS	6	0	12.0	19.9
19	OFFER	6	0	12.0	19.9
20	CAREER	6	0	12.0	19.9

Figure 2.2 Results generated from the BNC using the settings in Figure 2.1 for collocates of *beautiful* and *attractive*

In the activity illustrated in Figure 2.1, in a classroom context, the learner is focused on two words within a specific corpus and when they follow the process set out for them, they will generate the same results which they can then consider (Figure 2.2). This could be a whole class activity if the search were conducted using the classroom screen or whiteboard. On the other hand, this activity could work as an individual, paired or group task overseen by the teacher and discussed within the classroom setting so as to arrive at an understanding. Outside of the classroom setting, such an activity could be used in self-study. Indeed, it is important to note that Poole's 2018 book is also designed for self-study so learners could choose to undertake this search in an independent self-directed way.

There are many other possible freer uses of a corpus in the classroom or in self-directed learning. For instance, learners might use *any* corpus of their choice and report back on *any* interesting insights that they may have gained about how language is used. This kind of activity is at the opposite end of the Mukherjee's (2006) cline and aligns with the free-range view of discovery learning that we find in Bernardini's (2004) notion of the serendipitous use of corpora (see also Kennedy & Miceli, 2017). Criticisms of the free-range approach include its unanchored nature relative to language syllabi (*where does it fit with syllabi?*), the lack of control over the learning outcome and the teacher's loss of control (*how can the teacher control what the students are using the corpus for?*) (see Bernardini, 2004; Mukherjee, 2006; Kennedy & Miceli, 2017). However, fostering the skills for and motivation to conduct free-range self-directed exploration of a corpus is a goal that we value in our teaching.

Successfully finding out about language in this very free approach may lead to more learning. It may foster more sustained and embedded use of corpora as a reference tool for learners' independent use.

When comparing polar pedagogical positions such as those discussed above, it is useful to consider what sets them apart. Within the DDL narrative, we find this polarity expressed in terms of the counterpoints of *control* and *freedom* but perhaps this is too simplistic (see Figure 2.3).

In reality, teacher *control* and student *freedom* are just part of the picture and it is of course just a binary view between two points. When we unpack this theoretically,

Figure 2.3 The cline from *teacher control* to *student freedom*

we can plot a different cline – from *teacher mediated learning* to *student mediated learning*. In the middle of this, we can also consider *peer-mediated learning*.

Let us consider this cline in other ways (see Figure 2.4). What factors should also be taken into account? First, let us consider level of proficiency of the learners as a variable. How is the cline of *teacher control – student freedom* affected by the level of learner competence? Beginner level learners need more support or scaffolding while advanced learners should ideally be able to move more and more towards self-regulated corpus use once they are proficient in the use of a given tool or interface. A teacher can mediate corpus use at beginner and elementary level by choosing the teaching point, the data used (perhaps by taking examples from the corpus for the learners and using them on handouts or slides). What is often ignored as a variable is the level of technical proficiency and confidence of the teacher in terms of the use of corpora. This is a crucial variable here also. The teacher, at lower levels, can control what the learner does with the language from the corpus (i.e. the processes) so that it is instructionally structured and sequenced. However, being able to do this means being confident in one's methodology, proficient in the use of DDL and clear in one's lesson structure and planning. Teachers can also promote peer-to-peer learning by setting up collaborative tasks where the learning is enhanced through the co-construction of knowledge (for example, learners could be asked to undertake the task in Figure 2.1 in pairs and to prepare a mini-presentation on it). This more hands-off approach by a teacher takes no less skill and planning than the former example scenario.

The first call in this paper therefore is for a more fine-grained articulation of the theoretical underpinning of any DDL intervention that is being researched across a cline that is not based on the binaries of *teacher control* and *student freedom* but rather considers the theories of learning that underpin such binaries as well as many other variables that are important to how DDL is manifested. In other words, we need to intersect more with theories of learning and explore differing ontological positions within our research questions. If we take the broad parameters within Figure 2.4, we can explore many questions in detail that will draw out connections

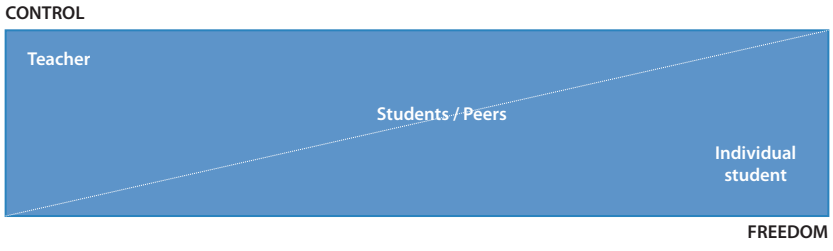

Figure 2.4 A more fine-grained model of control versus freedom in DDL

between how DDL is deployed pedagogically and whether it is successful within a given set of variables (e.g. age of students, proficiency level, cultural background, level of specialism and relevance of the corpora being explored, etc.).

On the left-hand side of this cline, we could plot teacher curation and control of *inter alia*:

- the teaching point: what language point do the learners focus on?
- the nature of the task: what the learners do?
- the nature of the corpus and data: what data is used (general, specialist, etc.)?; is it through a handout of teacher-mediated examples or a specific corpus that all students use?
- the degree to which this learning is scaffolded: what pre-teaching is done before the corpus task?; what help and prompting is given while students conduct the task?

In the middle of the cline, we can plot peer-to-peer learning:

- peer-to-peer construction of knowledge relative to the degree of teacher curation of teaching point, data and task.
- peer-to-peer scaffolding of knowledge relative to the degree of help and prompting given by the teacher during the task.

At the right-hand side of the cline, we can plot the type of free-range learning that is captured in what Bernardini refers to as "serendipitous corpus browsing" (2004, p. 22), where an individual student can:

- grapple with data (of their choice) to discover recurring and meaningful language patterns.
- self-regulate paths and processes of learning by virtue of competence in the use of DDL tools and through self-awareness and learner autonomy.

Figures 2.3 and 2.4 offer us purely schematic models but let us consider why they are important for DDL research. If a classroom-based study of DDL clearly plots its pedagogical stance in these terms, then we can aggregate results more systematically in terms of the impact of classroom processes on learning outcomes. Conversely, if we know little about the pedagogical stance and the nature of a DDL intervention within a given study, we are treating all types of DDL interventions as pedagogically uniform and gain little insight into what works best and why. In an ideal world, we could accrue results across teaching scenarios from teacher-controlled set-ups; peer-to-peer engagements and fully autonomous individual learning routes in future meta-analyses. This would be invaluable information because, as we shall discuss below, knowing more about how recurring patterns of language are acquired

by learners is crucial to understanding and informing instructed SLA. It may also prove crucial in bringing the merits of DDL to a wider audience.

Second language acquisition and DDL

Following on from theories of learning discussed above, this chapter is also concerned with the need to make links between theories of second language acquisition and DDL so as to provide a mutually-beneficial scenario where DDL practices might inform SLA and vice versa. As discussed, for almost three decades, the research paths of DDL and SLA research have generally not intersected (Flowerdew, 2015; O'Keeffe, 2020). This is unfortunate given the potential of both areas to inform each other, especially in relation to the question of how the brain works, on a conscious and a subconscious level, in the process of acquiring a second language through attention and noticing (Schmidt, 1990, 2001) of recurring patterns of language. O'Keeffe (2020) offers a detailed discussion on the importance of having a better understanding of SLA within DDL research in terms of the ongoing cognitive debate on how learning happens, whether consciously, sub-consciously (or a combination of both). She argues that DDL needs to be more aware of how it fits within this debate and how it can contribute to it through a broader research gaze.

In this section, we focus on some important links between DDL and SLA. Specifically, we zone in on one area of SLA research that has strong resonances with DDL, namely usage-based (UB) theory. By way of background, the UB model emerged from first language acquisition (FLA) studies (Tomasello, 2003) and subsequently gained traction in SLA studies (Bybee, 2008; Ellis, 2012). Essentially, it holds that our knowledge of language comes from experience and use, within a meaning-rich context, "as part of a communicatively-rich human social environment" (Ellis & Larsen-Freeman, 2006, p. 577). This model offers a frequency-based account of acquisition in the sense that encountering patterns of language is the key determinant of acquisition (Ellis, 2012). Through frequent and meaningful encounters with patterns of phonology, syntax and discourse, regularities emerge for a child acquiring a first language and for a learner acquiring a second one, or as Ellis (2002, p. 144) notes: structural regularities "emerge from learners' lifetime analysis of the distributional characteristics of language input". In other words, by experiencing language, we notice patterns and related meanings and ultimately acquire them (see Pérez-Paredes et al., 2020). It is believed therefore that through exposure to and use of language, our cognitive mechanisms make sense of the frequencies and regularities of forms we experience. Some of the central factors within the UB model of acquisition are:

1. *frequency* (the amount of times a construction is experienced and used)
2. *recency* (the more recently we experience a construction, the stronger our memory of it)
3. *context* (a given context triggers an association and mental categorisation of a frequently experienced construction)

The concept of the 'construction' is at the core of UB. Consider this UB perspective on FLA in a scenario of a child interacting with a guardian/carer. Imagine a child hears (or 'experiences') the following language examples:

Mammy's gone.
Mammy's gone in the car.
All gone.
Where's daddy gone?
It's gone!

Children first begin to construct language by putting two single words together in holophrases (word combinations). For example, from the above patterns that a child experiences, they might construct the two word phrase *car gone* while pointing to the driveway where the family car is normally parked (see Ellis, 2003). The next stage we expect is the abstracting of grammatical patterns where grammatical slots are filled and expanded. Therefore, *car gone* might expand to:

The car's gone,
Where's the car gone?
The juice is gone,
Where's the Lego gone?

At this stage, we see a movement from a holophrase formula to an abstraction of a pattern and its meaning which can lead to more "low-scope" pattern formation (see Ellis, 2003). Essentially, we see a transition from learning about what words go together (based on the language that is experienced) to learning about patterns of complementation, collocation and colligation on a verb-by-verb basis, as more new language is experienced (see Pérez-Paredes et al., 2020). In this way, the mind acquires construction patterns of form and meaning.

Constructions vary in terms of their complexity but, as Wulff and Ellis (2018) tell us, the more often a speaker encounters a particular construction (or combination of constructions), the more *entrenched* it becomes. To say that a construction is entrenched means that it has become *automatized* as a routine chunk of language that is stored and activated by the language user as a whole, rather than "creatively" assembled on the spot (De Smet & Cuyckens, 2007, p. 188). This essentially means a unit of meaning has been subconsciously stored in the brain of the language user.

As language users we have, as Wulff and Ellis describe it, "a huge warehouse of constructions that vary in their degree of complexity and abstraction" (2018, p. 39).

Within this UB paradigm, both first and second language learning involves an associative process of *tallying*, from an individual's accruing experience of language, the probabilities of occurrence of form-function mappings (see Pérez-Paredes et al., 2020). This view has been widely explored in FLA studies using rich and dense empirical data collected in language corpora, where the language of children, plus their encounters with care-givers, offers a solid body of evidence for the usage-based view (Tomasello, 2003) and it is increasingly accepted as a model for SLA. However, there are some key considerations in the context of instructed SLA compared to FLA. To begin with, in the context of SLA, learners are not usually very young children and their first language patterns are already well-formed. They have already gone through usage-based cognitive processes in their first language acquisition. This can both help and hinder the SLA process as we know from contrastive studies (see Granger et al., 2015). As Pérez-Paredes et al. (2020) note, the second language learner already has a well-developed schematised repertoire for at least one language. This point thus carries both positives and negatives to the context of instructed SLA.

Connections between UB perspectives and DDL are obvious. DDL offers learners a type of "condensed exposure" (Gabrielatos, 2005, p. 10) that can aid lexical and pattern awareness. It can bring a type of intensification of language experience through the data. DDL strives from a pedagogical perspective to accelerate the learner's experience of and engagement with structural regularities. For proponents of DDL, UB theory will resonate very much given that it holds that cognitive mechanisms are triggered through experiencing language patterns (see Pérez-Paredes et al. 2020). Essentially, DDL is all about giving learners repeated and intense *experience with* forms, in patterns (of morphemes/lexis, syntax and meaning) and thus it is imperative that links between DDL and UB theory be further explored.

Central to DDL has been the notion of *grappling* with raw data (Cobb, 2005). As Cobb notes, the DDL paradigm offers a methodology to help perform this grappling through adaptive tools and methods. Here it is also argued that a better understanding of the UB model could lead to a very fruitful research seam that investigates more closely the interface between intensive language exposure through DDL and the cognitive processes that might best facilitate learning. Let us explore this empirically by way of illustration of the potential that is yet to be tapped. Here we focus in on just one point that has emerged from meta-studies, namely that DDL works best for learners who are more advanced (cf. Boulton & Cobb, 2017). Through a UB lens, we could speculate that learners from intermediate level upwards have already

gained from building on low-scope patterns in the L2 and they are thus equipped to build on the cognitive processes that have already been used to acquire their L1. In other words, we can hypothesise that grappling with patterns in DDL, whether consciously or sub-consciously, is not daunting to more advanced learners because they have already abstracted many patterns and have a critical level of understanding of these patterns from their earlier language experiences with both their L1 and L2. However, we should not write off lower level learners from the advantages of DDL. If we work with the insights from the UB model, then we may be better able to curate and mediate the learning process for lower level learners so that they can experience language patterns that are differentiated to their level.

UB-based micro-insights into patterns of learning and how this might inform DDL

Research into the types of constructions that potentially best accelerate the acquisition process, at given levels of proficiency, would be exciting. Emerging work, for example by Römer (2019), explores, *inter alia*, the first patterns of verb-argument constructions (VACs) acquired by L2 beginners' level students and how this repertoire develops across levels. Römer's work is theoretically set within the UB model but, as a corpus linguist, she is shedding important light on learner corpus data that is highly valuable to the DDL community. We will now build on some of the findings from Römer (2019) and explore them in terms of how they might inform DDL.

Using a corpus of EF exams, the Education First-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013), Römer (2019) looks at a sample of German L1 learners of English. Among the top 10 most frequent VACs in these beginner-level data, Römer lists, for example (2019, p. 275ff):

- *be* copula constructions: a nominal subject, followed by a form of copula *be*, followed by a nominal complement: e.g. *My name is Anna*.
- Nominal subject, followed by a form of copula *be*, followed by an adjectival complement: e.g. *I'm happy*.
- Verb followed by prepositional phrase starting with *in*: e.g. *They live in Cologne*.
- Existential *there*, followed by a form of copula *be*, followed by a nominal subject: e.g. *There are many things near my house*.

Römer notes that longer, more complex VAC patterns are rare or non-existent at A1. While of course these results are to be expected in general at A1 beginner level, this study gives us *post-hoc* micro-insights into the patterns that seem to have been acquired. Also, it clearly shows the types of construction patterns that are established and thus can be built upon if one wishes to use DDL. In other words, it has

the potential to offer us a prototype for calibration of materials or input for DDL that is informed by a theory of second language acquisition.

Within DDL research, we can look at Römer's work in another way. A DDL study could offer rich data to complement the exploration of the acquisition of patterns at different levels of proficiency. Römer (2019) shows an exponential growth in the types of VACs acquired from A1 to C1 and within this, for example, we see a shifting of the repertoire in terms of what patterns learners most use. For example, there is a clear movement towards greater complexity of patterns, especially in relation to both nominal and adjectival complementation. Such work also gives us insights into the repertoire of forms that are used. For example, Römer (2019) found that the A1 learners rely on a narrow range of verbs. In addition to the copular *be*, the top 10 patterns were limited to the verb lemmas: *live, have, meet, like* and *see* (within the top ten VAC combination patterns) whereas at B2 level, the top patterns included the verbs copular *be, have, think, let, apply, find, want, observe* in more complex patterns of use (e.g. [lemma *observe*] + direct object). This kind of information is useful for a teacher wanting to use DDL with beginner's level students because it offers evidence of the patterns that are core to this level and on which learners build in their acquisition process.

However, it brings home strongly the need for careful curation of data for lower level DDL work, as Allen (2009) has advocated. When we take one of the A1 level patterns from Römer (2019), for example, existential *there* + copula *be* + a nominal subject, and look at its typical use among A1 level learners in general using the A1 data from the 53 million word Cambridge Learner Corpus (CLC) and contrast that with the typical use of the same pattern in the British National Corpus (BNC) written component, we can see important differences, as illustrated in Table 2.1, showing the top 20 patterns from the CLC and BNC dataset.

The first point to be made from our very basic exploration here is that the pattern is established at A1 level, however its use is different to that in the BNC. Let us consider the differences:

The A1 exponents of this pattern generally display a literal use that is focused on description and quantification of physical objects; there is evidence of singular and plural use of the copular verb *be*; there is no evidence of the past form of *be*; there is no evidence of negation in the top 20 forms.

The BNC exponents show a 40% use of the negative pattern *There is no*, all of which are used non-literally and seem to act at a discourse level in emphasis; we see variation in the forms of *be*, with singular and plural forms in both present and past; the forms appear to be used non-literally (i.e. not used in the literal description of physical objects); quantification is evident in a number of patterns (*there are a number of; there is/are a lot of; there is a great deal*).

Table 2.1 Top 20 most frequent exponents of *there* + copula *be* + a nominal subject at A1 level in the CLC and in the BNC written corpus

	A1 CLC	**BNC Written**
1.	there is a concert in	there are a number of
2.	there are a lot of	there is no doubt that
3.	there is a concert.	there is no need to
4.	there were a lot of	there is no reason why
5.	there were all my friends	there is no evidence that
6.	there is a concert on	there is a need for
7.	there is a concert of	there is no reason to
8.	there is my house.	there are a lot of
9.	there 's a concert in	there is no need for
10.	there is a concert,	there is no such thing
11.	there were a lot of	there is more than one
12.	there was my family and	there was a lot of
13.	there was a lot of	there is a danger that
14.	there is a concert next	there is a great deal
15.	there is a concert of	there is a lot of
16.	there were all our friends	there is little doubt that
17.	there is a concert at	there were a number of
18.	there is a rock concert	there 's a lot of
19.	there is a concert near	there is no point in
20.	there was all my family	there is likely to be

Why are these differences pertinent to DDL? In a nutshell, if we were to use the BNC as a source of examples or as a corpus for direct use with A1 level learners, the most frequent language that they experience would be far removed from the stage at which they are at in terms of abstracting this pattern. The evidence from the A1 data in Table 2.1 shows us that learners have established this form but only in the singular and plural present forms (*is/are*) and only in the literal meaning to describe and quantify the physical world. Beginner-level learners, faced with multiple instances of this pattern (in a native-speaker corpus like the BNC) in its multiplicity of exponents, singular and plural, past and present forms, in largely figurative senses, will obviously be utterly lost.

Interestingly, if we look at B2 and C2 level learner data in terms of this same form (in the CLC), we find that there is steady progression towards the figurative use of the pattern across its various forms (singular and plural; present and past). Table 2.2 illustrates the top 20 forms at A2, B2 and C2 level within the CLC and compares them with the BNC. The shaded cells show non-literal use of the form:

Table 2.2 Top 20 most frequent exponents of *there* + copula *be* + a nominal subject at A1, B2 and C2 level in the CLC and in the BNC written corpus (shading marks formulaic patterns with more figurative meanings)

	A1	B2	C2	BNC Written
1.	there is a concert in	there are a lot of	there are a lot of	there are a number of
2.	there are a lot of	there is a lot of	there is no doubt that	there is no doubt that
3.	there is a concert.	there were a lot of	there are many people who	there is no need to
4.	there were a lot of	there is a new direct	there was no money left	there is no reason why
5.	there were all my friends	there were over 5,000 people	there is a lot of	there is no evidence that
6.	there is a concert on	there is a new collection	there are some people who	there is a need for
7.	there is a concert of	there is a new shop	there is no need to	there is no reason to
8.	there is my house.	there is no doubt that	there is no point in	there are a lot of
9.	there's a concert in	there is no need to	there is a lack of	there is no need for
10.	there is a concert,	there are many things to	there are more and more	there is no such thing
11.	there were a lot of	there were no discounts available	there is no need for	there is more than one
12.	there was my family and	there are advantages and disadvantages	there were a lot of	there was a lot of
13.	there was a lot of	there are some things that	there are people who are	there is a danger that
14.	there is a concert next	there is a lack of	there are people who do	there is a great deal
15.	there is a concert of	there are a few things	there are a number of	there is a lot of
16.	there were all our friends	there are too many cars	There are, of course	there is little doubt that
17.	there is a concert at	there was a lot of	There are, however,	there were a number of
18.	there is a rock concert	there is a need for	there are people who believe	there 's a lot of
19.	there is a concert near	there was a different actor	there is a number of	there is no point in
20.	there was all my family	there 's a lot of	there is a wide range	there is likely to be

Here we see a progression in the acquisition process. Learners' most frequent uses of this pattern, which they have established at A1, show evidence of abstraction where increasingly they are able to use the pattern both literally and figuratively. They are increasingly showing that they have abstracted its discourse function where it can be used for emphasis through its negated form. This probably explains why more advanced learners have greater success with DDL (illustrated with an example from C2 level in the CLC):

> *As far as I am concerned, I dwell in the center of the town. In my way of thinking* ***there is no doubt that*** *there are considerable problems related to overcrowdedness and the low standard of accomodation.* [C2: Greek; CPE; 2009]

Within this small vignette of corpus analysis, there is considerable food for thought for DDL design. Data-driven learning should not just be about flooding learners with any data. We need to engage more with the emerging UB corpus-based research that can tell us much about how best to curate data for lower level learners. We can only then really address whether DDL is suitable for lower learners. For sure, as we have illustrated here, using data from a native-speaker corpus with A1 level learners will not work because the data is not differentiated for their level because they are still establishing the variations of form, meaning and use of the pattern. If data is selected for them at an appropriate level, with very specific learning outcomes, then it is hypothesised that it may accelerate their learning. A micro example of this might be to focus in on the quantification pattern discussed above. If learners were to be exposed to patterns of quantification, they could first work on variation of form based on corpus information:

There's	a lot +
There is/are	a lot +
There was/were	a lot +
	a lot +
	many
	over +
	some +
	more than +
	a great deal +
	a number of +

This would ultimately allow for learners to work towards patterns relating to quantification that are non-literal and are formulaic, such as those illustrated in Table 2.2, e.g. *there is no doubt that, there is no point in; there is a lack of*, etc. However, the differentiated (for level) curation of form is only part of the picture. Let us return

to the three central factors in acquisition within the UB model as detailed above: (1) *frequency*, (2) *recency* and (3) *context*. While frequency and recency can be attended to within our current understanding of DDL, *context* is a big challenge (i.e. offering the meaningful context to trigger an association and mental categorisation of a frequently experienced construction).

As discussed, the UB model of language acquisition centres on meaningful input and holds that the observation of frequency and patterning of form and meaning is core to language learning. Acquisition of constructions is connected with exposure to meaningful form-function relations upon which a learner builds a language system. However, the list of quantification patterns above is not presented in any meaningful context. Were they to be offered as full sentence examples drawn from a corpus, they still would not offer the rich meaningful experience from which we abstract meaning in naturally-occurring acquisition. This is a major challenge for using DDL in the optimum way for acceleration of acquisition and it boils down to addressing how we can better encode meaning. In other words, how can the intensification of input also be meaningful so as to aid form-meaning mapping? The answer may well be in the development of better interfaces, ones that are richer in their use of multi-media where sound, pictures, video are embedded in the overall experience. This is a point made clear by Meunier (2020, p. 19):

> there is also room for more creativity in the DDL tasks that could be proposed to younger learners, especially keeping in mind the affordances of current digital tools (multimodality, gaming options, easy access, intuitive use, etc.).

She points out that concordances are not the only possible "triggers of frequency effects and form-meaning mappings in focus on form activities". In natural first and second language acquisition, form-frequency mapping is a multi-modal experience (usually with audio, visual, verbal, non-verbal, prosodic information). A concordance line is offering only a fraction of the real experience of written language. One suggestion or example Meunier offers is the application *Playphraseme* (see https://www.playphrase.me). This tool allows users to search for phrases. By inserting a search word or string, the user is presented with patterns that are linked to video clips from films and series where the phrase is used. For example, a beginner-level learner can insert the phrase *there's a lot* and immediately short video snippets plus sub-titles with the search phrase highlighted will displayed, along with the video clip and sound. This is an example of an infinitely more meaning rich way of doing DDL that incorporates an experience of language that is multi-modal and closer to a real experience in terms of accelerating the learning process.

Conclusion

This paper has argued that while we have gained so much insight from existing work on DDL, we have largely not engaged with the theoretical underpinnings of teaching, learning and acquisition and this may be impeding the mainstreaming of DDL. Through greater quantification of the pedagogical stance and beliefs of the teacher who uses DDL, those who research it can find out more about which approach works best in which contexts.

Greater connection with theories of SLA, especially the UB perspective, will mean that we can be part of inquiry processes that may lead to micro-understandings of the connection between teaching and learning and form-meaning mapping and abstracting. In so doing, DDL can be part of a two-way process where the findings of DDL also enrich instructed SLA models and perspectives.

Without doubt, a rich seam of research on DDL exists and it has shown that this is a fruitful approach but, to evolve, there is a need to open up new debates and motifs within our research community. A broader DDL research narrative can lead us to an enhanced rationale for and understanding of the benefits of using DDL in the classroom, one which is based on and interlinked with theories of learning and second language acquisition.

Some of the questions that need to be addressed in greater depth include those around how learning takes place in DDL. We need to understand more about:

- how the exploration of regularities of patterning can intersect with meaning; the degree to which this might be relative to individual learners, levels of proficiency and instructional design, for example.
- how degrees of pedagogical control and freedom, mediation and self-regulation interplay with acquisition of form and meaning, and again how this might be relative to variables such as level of proficiency, and so on. The less we know about the pedagogy that underpins a DDL study, the less we learn about what works.

In this chapter, we have also focused on learner data to explore in a very rudimentary way the exponents of one verb-argument construction in one learner corpus. There is so much more work to be doing in terms of exploring constructions across levels of proficiency so as to better understand acquisition and interlanguage. DDL researchers can work within this process by testing the *learnability* of constructions and by exploring the interface of learning (e.g. eye-tracking, multi-modal input enhancement, etc.). Such explorations have a lot of potential for SLA experimentation.

In summary, an intersection of DDL, theories of learning and second language acquisition theory is the key to bringing DDL to a more mainstream audience. The essential message of this paper is that we know a lot about the D for *Data* in DDL but we need to think much more about the L for *Learning*. We need to think about

the nature of this learning; we need to think about the connection between *how* we teach and its impact on learning, and we need to think about *what* we teach in a more differentiated way, relative to the stage of acquisition of our learners. To put it starkly, we need to engage with theories of learning and models of second language acquisition if we are to move beyond driving text-based data at learners in the hope that some of it will stick in their subconscious store.

References

Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal*, 63(1), 23–32.

Bernardini, S. (2004). Corpora in the classroom. An overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15–36). John Benjamins. https://doi.org/10.1075/scl.12.05ber

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, 60(3), 534–572.

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

Bybee, J. L. (2008). Usage-based grammar and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 216–236). Routledge.

Cobb, T. (2005). Constructivism, applied linguistics, and language education. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (Vol. 3, 2nd ed., pp. 5–88). Elsevier.

De Smet, H., & Cuyckens, H. (2007). Diachronic aspects of complementation: Constructions, entrenchment and the matching-problem. In C. Cain & G. Russom (Eds.), *Studies in the history of the English language III: Managing Chaos: Strategies for identifying change in English* (pp. 1–37). De Gruyter Mouton.

Ellis, N. C. (2002). Frequency effects in language processing a review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143–188.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 33–68). Blackwell.

Ellis, N. C. (2012). Frequency effects. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. 260–265). Routledge.

Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics. Introduction to the special issue. *Applied Linguistics*, 27(4), 558–589.

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15–36). John Benjamins. https://doi.org/10.1075/scl.69.02flo

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *Teaching English as a Second or Foreign Language*, 8(4), 1–34.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In *Proceedings of the Conference ICT for Language Learning 2013*, R. T. Miller (Ed.), Florence, Italy.

Granger, S., Gilquin, G., & Meunier, F. (Eds.) (2015). *Cambridge handbook of learner corpus research*. Cambridge University Press.

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

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 33–44). John Benjamins. https://doi.org/10.1075/scl.33.05joh

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 293–313). Cambridge University Press.

Johns, T., & King, P. (1991). Classroom concordancing. *English Language Research Journal* 4, 17–25.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology*, 5(3), 77–90.

Kennedy, C., & Miceli, T. (2017). Cultivating effective corpus use by language learners. *Computer Assisted Language Learning*, 30(1–2), 91–114.

Lantolf, J. P., & Ahmed, M. K. (1989). Psycholinguistic perspectives on interlanguage variation: A Vygotskyan analysis. In S. M. Gass et al. (Eds.), *Variation in second language acquisition: Psycholinguistic issues* (pp. 93–108). Multilingual Matters.

Lantolf, J. P., & Appel, G. (Eds.) (1994). *Vygotskian approaches to second language research*. Ablex.

Lee, H., Warschauer, M., & Lee, J. H. (2019). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*, 40(5), 721–753.

Meunier, F. (2020). A case for constructive alignment in DDL: Rethinking outcomes, practices and assessment in (data-driven) language learning. In P. Crosthwaite (Ed.), *Data-driven learning for the next generation. Corpora and DDL for pre-tertiary learners* (pp. 13–30). Routledge.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art – and beyond. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy* (pp. 5–24). Peter Lang.

O'Keeffe, A. (2020). Data-driven learning – A call for a broader research gaze. *Language Teaching*, 1–14. https://doi.org/10.1017/S0261444820000245

O'Keeffe, A., M. J. McCarthy, & R. Carter. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

O'Sullivan, Í. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, 19(3), 269–286.

Pérez-Paredes, P. (2010). Corpus linguistics and language education in perspective: Appropriation and the possibilities scenario. In T. Harris & M. Moreno Jaén (Eds.), *Corpus linguistics in language teaching* (pp. 53–73). Peter Lang.

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

Pérez-Paredes, P., Mark, G., & O'Keeffe, A. (2020). *The impact of usage-based approaches on second language learning and teaching*. Cambridge University Press.

Poole, R. (2018). *A guide to using corpora for English language learners*. Edinburgh University Press.

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

Römer, U. (2019). A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, 24(3), 268–290.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *ReCALL*, 26(2), 184–201.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tribble, C. (2008). From corpus to classroom: Language use and language teaching. *ELT Journal*, 62, 213–216.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford University Press.

Wulff, S., & Ellis, N. C. (2018). Usage-based approaches to SLA. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language: The state of the science across its subfields* (pp. 37–56). John Benjamins. https://doi.org/10.1075/sibil.54.03wul