Applied Corpus Linguistics 2 (2022) 100028

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Principled pattern curation to guide data-driven learning design

Anne O'Keeffe [a,*], Geraldine Mark [b]

[a] Mary Immaculate College, University of Limerick, Ireland
[b] Cardiff University, United Kingdom

ARTICLE INFO

Keywords:
Data-driven learning
Learner corpus research
Usage based acquisition
Complexity
Formulaicity

ABSTRACT

Insights from corpus linguistics (CL) have informed language learning and materials design, among many other areas. An important nexus between CL and language learning is the use of Data-Driven Learning (DDL), which draws on the use of corpus data in the classroom and which brings opportunities for inductive language discovery.

Within the ethos of DDL, learners are encouraged to discover patterns of language and, in so doing, foster more complex cognitive processes such as making inferences. While many studies on DDL concur on the success of this approach, it is still perceived as a marginal practice. Its success so far has been largely limited to intermediate to advanced level learners in higher education settings (Boulton and Cobb 2017). This paper aims to offer guiding principles for how DDL might have wider application across all levels (not just at Intermediate and above) and to set out exemplars for their application at different levels of proficiency. Based on insights from second language acquisition (SLA) and learner corpus research (LCR), the focus of this paper will be on identifying principles for the curation of language patterns that are differentiated for stage of learning. In particular, we are keen to build on recent and important work which looks at SLA through the lens of the usage-based (UB) models (that is, models that view language as being acquired through the use of and exposure to language).

1. Introduction

Over the last three decades, Data-Driven Learning (DDL) has been widely championed through scholarship by those of us who see the exciting opportunities that it can bring to the language learner in terms of inductive language discovery. This method was heralded as a means of turning linguists' analytical procedures into a pedagogically-relevant tool to increase both learners' awareness of and sensitivity to patterns of language, while also enhancing language learning strategies (Pérez-Paredes, 2010). It has been widely claimed that the inductive processes of engagement with a corpus can foster more complex cognitive processes such as inference and hypothesis-formation (O'Sullivan, 2007). A number of surveys and meta-analyses concur on the general success of this approach (see Boulton and Cobb 2017; Vyatkina and Boulton 2017; Lee et al., 2019, among others) but as Boulton (2017: 1) notes, DDL is still a "marginal practice". Its success seems to be limited to intermediate to advanced level learners in higher education settings (Boulton and Cobb 2017). This paper aims to offer guiding principles for how DDL might have wider application across all levels of proficiency. The focus of this paper is to take stock of and aggregate findings from second language acquisition (SLA), learner corpus research (LCR) and DDL so as to propose and inform a framework to enhance data-driven design. The framework will identify research-based principles for the curation of language patterns that are differentiated for stages of learning, in line with the Common European Framework of Reference (CEFR). In particular, we are keen to build on recent and important work which looks at SLA through the lens of usage-based (UB) models (those that view language as being acquired through use and exposure). This work underscores the importance of frequency of exposure to natural language in the language learning process, which DDL can offer (see Meunier, 2020; Pérez-Paredes et al., 2020; O'Keeffe, 2021a). Our survey of SLA and LCR (Sections 2 and 3) will be followed by our proposed framework and principles for the enhanced curation of patterns in DDL. This will then be exemplified through task design case studies.

2. Second language acquisition (SLA) research and DDL

Frequency of encounter and occurrence is key to both UB models of acquisition (Ellis, 2012) and to DDL. In DDL, there is a focus on guiding learners towards regularities so as to become aware of generalisations in patterns of form and meaning (O'Keeffe, 2021a). It can be argued that DDL can bring an acceleration of language (frequency) experience to the learner, through a type of 'input flooding' (after Sharwood Smith, 1993). DDL can offer learners a type of 'condensed exposure' (Gabrielatos, 2005: 10) that can aid lexical and pattern awareness and can drive a type of intensification of the cognitive process

* Corresponding author.
E-mail addresses: anne.okeeffe@mic.ul.ie (A. O'Keeffe), markg2@cardiff.ac.uk (G. Mark).

https://doi.org/10.1016/j.acorp.2022.100028
Received 27 July 2022; Accepted 4 August 2022
Available online 6 August 2022

HOLOPHRASAL SEQUENCES	EXPANDED SLOTS AND FRAMES	FULLY ABSTRACTED PATTERNS
	→	
I'd like	**pronoun + modal + base verb** *She + 'll + want* *They + can + go*	*I must admit* *It may sound (odd) but*
I went to the cinema	**pronoun + past simple + prep + det + noun** *I + went + to + the + cinema* *We + walked + across + the + street* *They + came + to + our + house* *She + ran + up + the stairs*	*I came to the conclusion* *It came as no surprise*
The centre of the city	**det + noun + prep + det + noun** *the + middle + of + the + room* *a + crack + in + the + ceiling* *The + aim +. of + the + proposal* *The + reason + for + this + request*	*a drop in the ocean* *(It's not) the end of the world* *the centre of my universe* *a gap in the market*

Fig. 1. Process of language acquisition within a UB model, with examples.

through 'grappling' with patterns (O'Keeffe, 2021b). However, some pedagogically central questions have yet to be answered: 'What DDL input will best promote language acquisition and development?' and 'Does this depend on the level of the learner (i.e. their stage of development)?'. Related to this, the traditional ethos of DDL is to promote independent inductive discovery through free corpus foraging, however, this requires at least an intermediate level of competency. O'Keeffe (2021b) makes the case for a more mediated process in DDL where the teacher takes a greater role in curation and task differentiation (by level). This is especially necessary if learners at lower levels are to use DDL successfully. If this is to happen, there is a pressing need for guiding principles for the curation of patterns and DDL task design so as to optimise learners' 'condensed exposure' to language.

2.1. Aligning DDL with acquisition processes

In the traditional constructivist definition of DDL, the curation of patterns from the corpus is ideally driven by a student's curiosity, leading them through discovery and induction from form(s) to meaning(s). The ideal user of DDL is motivated to investigate and abstract meanings from patterns of language use in the corpus and ultimately to store these patterns so that they can form part of their repertoire of language, which can be expanded over time (O'Keeffe, 2021a). DDL intervention studies show that this technique carries pedagogical merit and is worth the technical effort on the part of the teacher and the student (e.g. Cobb and Boulton 2015; Boulton and Cobb 2017; Vyatkina and Boulton 2017; Lee et al., 2019). However, O'Keeffe (2021b) argues that very little thought has gone into the rationale for *why* repeated encounters with patterns of language might be a good idea from a theoretical perspective of SLA. Many have called for connections to be made between DDL and SLA (Flowerdew, 2015; Johansson, 2009) especially via a UB model of acquisition which is seen to align well with this approach (Ellis, 2012; O'Keeffe, 2021a, 2021b; Pérez-Paredes et al., 2020; Römer, 2019). UB evidence suggests that the process of learning an additional language, as with a first language, involves intentional pattern finding which develops along a cline from basic formula (word combinations) to slot and frame sequences to fully abstracted constructions (Ellis, 2003; Pérez-Paredes et al., 2020). In other words, second language learners typically move from a repertoire of fixed holophrasal sequences at low levels to an expanded slot and frame system to fully abstracted (often figurative) patterns as illustrated in Fig. 1.

A core tenet of the UB model is that our knowledge of language comes from experiencing and using it as part of a communicatively-rich human social environment (Ellis and Larsen-Freeman 2006). As language develops, we see a transition from learning about what words go together to learning about patterns of complementation, collocation and colligation, as more new language is experienced (see Pérez-Paredes et al., 2020). In this way, the mind acquires 'constructions' routinised patterns of form and meaning (Langacker, 1987). Constructions vary in terms of their complexity ranging from morphemes, e.g. affixes like *in-* in *incredible*, to words to phrases to more abstract syntactic frames, such as the ditransitive construction, *give something to someone*, carrying a meaning related to 'transfer'. UB theorists hold that learners are (subconsciously) aware of the frequency of occurrence of constructions and the more often they encounter 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 subconsciously stored and activated by the language user as a whole, rather than 'creatively' assembled on the spot (De Smet and Cuyckens, 2007: 188). As language users, we have, to quote Wulff and Ellis, "a huge warehouse of constructions that vary in their degree of complexity and abstraction" (Wulff and Ellis, 2018: 39).

2.2. Frequency, categories and prototypes

One of the most important insights that CL has revealed about language usage lies in the distributions of frequencies of different linguistic features. In different contexts, some morphemes, words, phrases, chunks, sequences, constructions occur more frequently than others, simply because they are more useful and therefore more used than others. Frequency in language is a natural phenomenon which can be described in terms of Zipf's law (Zipf, 1935), a power law which describes the relationship between the frequency of units of language and their frequency rank (Piantadosi 2014) where the frequency of a linguistic unit is inversely proportionate to its rank. In naturally-occurring language the first, most frequently occurring, word occurs twice as often as the second most frequent word and three times as often as the

third most frequent word, etc. For example, in corpora of naturally-occurring spoken English, *the* is typically the most frequently occurring word, and it occurs approximately twice as often as the second highest ranking word *and*, and three times as often as the third highest ranking word, etc. This relationship can be seen not only across individual words, but also, across *other units of language*, other than words, for example constructions (Ninio, 2005, Ellis et al., 2016) and users of language are subconsciously aware of this phenomenon, through statistical learning. As we encounter new language, we categorise it, matching it to what we already have in our subconscious store. For example, we know that *oranges, lemons, limes, grapefruit* are all part of the category of *citrus fruit* and when we come across an unusual fruit like *a kumquat* for the first time we are likely to put it into the *citrus fruit* category because it displays prototypical characteristics. We also have a sense through statistical tallying and categorisation that *orange* is likely to be the most frequently occurring, a prototype for the category. Extensive research has also shown that constructions have prototypes; for example, in verb argument constructions (VACs), the types of verbs occupying the verb slot of any construction share characteristics of the prototypical meaning and also have a Zipfian distribution (Goldberg, 2006; Ellis et al., 2016). For each VAC, there is one verb, which Ninio terms 'pathbreaking' (1999), which takes the largest share of the distribution and which is prototypical of the meaning of the construction. For example, in the VL (verb locative) construction, movement to place, *go* is the prototype verb, followed by *come*; in the VOO construction (verb + object + object), *give* is the prototype, followed by *send*. When learners come across subsequent verbs found in the same syntactic contexts, or slots, in the input, they draw on the prototype from which to infer meaning (Römer and Garner 2019; see also Section 3 below). These prototypes are "the hubs in the construction's semantic network" (Ellis and Ogden 2017: 609). As we acquire these form-meaning mappings, we learn to categorise. As we build our linguistic repertoire, we learn to match the new words, phrases, structures that we come across for the first time against what we have already encountered and categorised. It is important to emphasise here that these categories and prototypes exist at all levels of abstraction, e.g. affixes like *in-* in *incredible*, to words to phrases to more abstract syntactic frames. How then can we apply this understanding of language development to DDL, to identify and accelerate which patterns to point learners to?

2.3. The importance of curation and mediation

We argue that the UB model can help us understand why meta-studies such as Boulton and Cobb (2017) find that more advanced learners are suited to grappling with language patterns in DDL. We posit that it is because learners at these levels have already abstracted many patterns and have attained a critical level of understanding of these patterns in terms of mapping their forms and meanings[1]. Through a UB lens, therefore, it can be speculated 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. Conversely, we hypothesise, guided by analysis of learner corpora, that learners below intermediate level have not abstracted enough patterns to cope with typical concordance lines usually drawn from native speaker corpora in DDL (see Section 4.1). However, we argue that learners who are at lower levels of proficiency should not be excluded from the advantages of DDL and, as a result, we underscore the need for careful and principled design in terms of how DDL is used at these lower levels (e.g. CEFR levels A1 to B1) so as to structure the process of acquisition based on a UB-based understanding of language acquisition. We show that an understanding of development in learner language through the use of learner corpora can inform this. As we shall discuss, the insights from the UB model and our understanding of how language develops may offer guidance in the curation and mediation of data and tasks for lower-level learners so that they can experience language patterns that are differentiated to their level (see Sections 3 and 4 where we develop this point).

3. Learner corpus research (LCR) and how it might inform DDL

3.1. Defining proficiency in learner corpora

Learner corpora offer an important testbed for identifying how we might better curate patterns across developing levels of proficiency when using DDL but hitherto their potential to inform the curation and task design process has not been fully realised. To evolve a framework to guide the differentiated curation of patterns for DDL, we hold that it is essential to take stock of the key findings from LCR and how they align with SLA research, particularly with UB studies discussed in Section 2. By engaging in this process, we will propose a workable framework as we outline in Sections 3 and 4.

We argue that, given the key role of frequency in acquisition as discussed above within the UB model, analysis of learner corpora that are structured by level of proficiency can help us look at learners' representations of language use which reflect where learners are in terms of working out the "probabilities of occurrences" of form-meaning mapping (Ellis, 2012: 196). Over recent years, the sampling of learner data by level has emerged as a more reliable variable for the exploration of language acquisition. This shift moves from using schooling year or age as proxies for language competence (Meunier, 2015) to attested performance levels usually based on standardised examinations (Green, 2010). As noted by Tono and Díez-Bedmar (2014: 165) and Forsberg Lundell (2021), the use of the Common European Framework of Reference (CEFR) levels of proficiency is emerging as a standardising measure, for example in the design and compilation of new corpora, and particularly in Europe, as a means to compare like with like in learner corpus data (see Harrison and Barker 2015; Hawkins and Buttery 2010; Hawkins and Filipović 2012; O'Keeffe and Mark 2017; Thewissen, 2013). For recent examples see the Trinity Lancaster Corpus (TLC) (Gablasova et al. 2019), and the EF-Cambridge Open Language Database (EFCAMDAT) (Alexopoulou et al., 2015).

3.2. Descriptions of development using learner corpora

To conduct research into learner competence by level of proficiency, large corpora that are calibrated to the CEFR are required and these are usually, though not exclusively, linked to exam corpora. For example, Hawkins and Filipovic (2012), and Hawkins and Buttery (2010) used the 55-million word Cambridge Learner Corpus (CLC), a corpus based on Cambridge exams across more than 200 countries and 140 L1 backgrounds across the six levels of the CEFR. In their study, they identified a series of 'criterial features', properties that were seen to characterise and point to L2 proficiency, at each of the CEFR levels. Murakami and Alexopoulou (2016) also used the CLC to evaluate the long-held view of a universal order of acquisition for English morphemes (Dulay and Burt 1973). They concluded that there was a strong L1 influence in the accuracy of the morphemes, which affected different morphemes in different ways, and refuted the universal order of acquisition theory.

In another study using the CLC, O'Keeffe and Mark (2017) developed the English Grammar Profile (EGP), a generic profile of learner use of multiple grammatical features (descriptors), traditionally covered in English language teaching classroom contexts, across six proficiency levels. In this pseudo-longitudinal study, using a criteria-based methodology, they observed development as an expanding repertoire of lexis, patterns and functions, as well as pragmatic competence. It was noted that as proficiency increased, learners put syntactic patterns, previously acquired at lower levels, to multiple uses. To do this they draw

[1] We acknowledge that there are other important pedagogical considerations in relation to the challenges faced by learners are lower levels in DDL including task complexity.

on an expanding lexical repertoire, while displaying a greater awareness of the collocational and colligational limitations of a given pattern, as well as a growing understanding of specialised pragmatic meanings (see O'Keeffe and Mark 2017). Of relevance to this paper, the output of the EGP is a database of 1,222 descriptors of grammatical competence across the six levels of the CEFR. This serves as a generic description of what learners can do with grammar at each level of proficiency. In a parallel project, Capel (2010) developed the English Vocabulary Profile (EVP), describing the words and phrases used by learners at each CEFR level. Both the EGP and EVP resources have applications for the curation of patterns for DDL which we will discuss further in section 4.

Using another large pseudo-longitudinal corpus, the 33-million word EFCAMDAT, Alexopoulou et al. (2015) examined relative clauses to demonstrate how large datasets can be used to study developmental trajectories across proficiency levels. Their findings indicate L1 effects and show how different types of relative clauses increase with proficiency. Thewissen (2013) looked longitudinally and contrastively at sample lexical and grammatical items, tracking learner development across four proficiency levels (B1, B2, C1, C2) specifically in relation to accuracy. She tracked the developmental pathways of error types in an error-tagged sample of the ICLE and observed strong progress (in terms of error decrease) between B1 and B2 levels. She observed a plateauing of progress in relation to errors between B2 and C2 levels which she posits may "hide qualitative development" (Thewissen, 2013: 87). This is in line with O'Keeffe and Mark (2017) discussed above.

In two studies, Pérez-Paredes and Díez-Bedmar (2019) and Díez-Bedmar and Pérez-Paredes (2020) use the Spanish learner component of the International Corpus of Crosslinguistic Interlanguage (ICCI) comprising 17,034 tokens (see Tono and Díez-Bedmar 2014). They use a combination of methods to measure syntactic complexity, across a range of age groups (grades 8 to 12). Both studies point to the analysis of complexity of the noun phrase as being "of great interest … in terms of identifying development milestones in language acquisition" (Pérez-Paredes and Díez-Bedmar 2019: 101). We return to this in Section 3.4 in the context of phrasal and clausal development. First, we review an important body of work on lexical bundles; these studies also bring insight to the importance of the noun phrase in language development.

3.3. Lexical bundles and development

Lexical bundle studies (of sequences of three or more words that co-occur frequently in a particular register) have also noted the importance of the noun phrase in development (Biber et al. 1999). Chen and Baker (2010) compared their form and function in L1 and L2. They found that while the structural and functional features of lexical bundles in both datasets were similar, learners had a tendency to use more verb-based bundles than L1 expert writers who demonstrated a wider range of noun-based structures. In a subsequent study, Chen and Baker (2016) took a developmental perspective, benchmarking L1 Chinese data from the Longman Learner Corpus (LLC) to CEFR proficiency levels. They examined four-word lexical bundles across B1, B2 and C1 level data. Lower-level learners (B1) were found to use of verb-based bundles. These were closer to conversational bundles, reflecting functions of personal interaction and quantity. In contrast, higher levels learners used bundles more characteristic of academic prose, with a higher proportion of noun and preposition-based bundles, reflecting a more impersonal tone. At B2 level, learners start to become sensitive to the bundles that index differences in formality (Chen and Baker, 2016).

Vidakovic and Barker (2010) examined four-word lexical bundles in 100 written texts from the Cambridge Skills for Life data (part of the Cambridge exam suite), across proficiency levels A1 to C1 and found that higher proficiency levels used a wider range of bundles and with greater frequency than at lower levels. Their functional analysis showed an increase in recurrent stance-indicating and discourse-organising use as proficiency increased. Staples et al. (2013) also used exam data, the Test of English as a Foreign Language (TOEFL iBT), to look at lexical bundle frequency and usage across three proficiency levels (loosely described as low, medium and high). Across all levels, they found stance-indicating bundles were most prevalent, and these tended to reflect the immediate context and topics of the exam prompts. Additionally, they looked at variability of fixedness, degrees of formulaicity, within bundle slots. Unlike Vidakovic and Barker (2010), their results showed a decrease in frequency of fixed bundles at higher levels which they propose was linked to a lower-level reliance on bundles from the exam task prompt (Staples et al., 2013). This contributed evidence to support a developmental sequence in some aspects of formulaicity, as proposed by Ellis (2002) within a UB model, in which learners move from a heavy reliance on holistic patterning at lower levels to 'self-constructed' sequences (Ellis 2002: 145) as proficiency increased (Staples et al. 2013). This suggests a move from formula to a slot and frame system (see also Section 2). In this UB developmental model, there is also a further step of abstraction, in which formulaicity plays a key role, increasing with proficiency (Ellis et al., 2016). This observation is also corroborated by Lenko-Szymánska (2014) ICCI-based study of 3-gram lexical bundles, across six L1 backgrounds, from A1 to B2 levels. Aligning with many other findings hitherto discussed, she found that formulaicity increases with proficiency and that bundles containing verb fragments were used at lower levels whereas bundles containing noun and prepositional phrases were seen at higher levels of proficiency. These consistent findings about the reliance on verb phrases at earlier stages of learning giving way to the development of noun phrase complexity as well as an increase in formulaicity are important points for DDL design and we explore their implications for and application to DDL in Section 4. In the next section, we first look at important findings from LCR that relate to phrasal and clausal development.

3.4. Phrasal and clausal development

Biber and Gray (2011, 2016) highlight the phrase and 'compressed phrasal structure' as an equally important indication of grammatical complexity and development as clausal structure and dependence (Biber et al., 2020). Alongside the phrasal complexity, they point to the role of register awareness in the developmental process. As part of this process, compressed phrasal structure takes centre stage in development as learners become more aware of its importance in writing. Biber et al. (2011, 2020) offer five hypothesised stages of development which indicate a general trend towards a decreased use in dependent clause complexity and an increased use of phrasal complexity (from finite complement clauses to pre and post modified noun phrases). They call for descriptions of writing development that include frequently used devices that mark the phrasal compressions such as premodification of nouns with attributive adjectives, and prepositional phrases as post-modifiers (e.g. *increase in inflation rates*).

An important point for our proposed framework is that more attention is needed on not just continuous lexical bundle sequences, but also on discontinuous strings, variously referred to as collocational frameworks, lexical frames, phrase frames or p-frames (i.e. recurrent strings in which not all words are fixed e.g. *on the ?*). Gray and Biber (2013), looking at L1 data, note the need to examine discontinuous sequences in their own right as linguistic building blocks. Their findings reveal that the frames that appear most frequently in academic writing consist of function words (e.g. *in the ? of, the ? of the*) (Gray and Biber, 2013). We return to this point in Section 4 to show how corpus software can now facilitate DDL task design that focuses on high frequency discontinuous bundles.

In a very relevant study of L2 writing, across proficiency levels, Garner (2016) examines p-frames in the German subsection of the EFCAMDAT. The p-frames are classified both structural and functional (after Gray and Biber 2013 and Biber et al., 2004). Crucially for the purposes of this paper, Garner shows that more proficient learners introduce more variability into their frame usage, especially between B2 and C1

level. Garner concludes that lower level learners rely more on fixed type frames whereas higher level learners employ a greater range of phraseological items. Taking a UB perspective, Garner accounts for these results by proposing that higher level learners have had more exposure to English, across a wider variety of contexts and therefore would have encountered more p-frame exemplars, with the effect of "entrenching p-frames in the learners' linguistic inventories" (Garner 2016: 49).

UB studies that use corpus data to examine *verb argument constructions* (VACs) in L2 language are growing (Ellis et al., 2016; Römer et al., 2014; 2018). Römer and Garner (2019) investigate five VACs constructions, in the Trinity Lancaster Corpus Sample (TLCS), using an L1 Italian and Spanish subcorpora (c. 1 million words). Their study gains insight into development of verb construction knowledge, comparing the findings with L1 usage using the BNC as a benchmark. They observed:

- Strong consistency in the choice of lead verbs for each VAC, suggesting that learners at all levels are sensitive to frequency of usage and have an awareness of appropriate candidate verbs for the verb slots.
- As proficiency increased, it aligned more with the L1 data; the distribution of usage in the C1/C2 data, compared to distribution in the B1 data, was found to be closer to the BNC results. Also, the variety of verb forms for each VAC in the higher level learners B2 to C2 was seen to be closer to the L1 data than in the lower level learners.
- An overall development of VACs usage, in line with growing proficiency, moving from a small set of fixed patterns to a larger set of more varied patterning, with usage becoming increasingly predictable and more Zipfian.

In a related study of VAC usage in L1 German learners in a 6 million word sub-corpus of EFCAMDAT, Römer (2019) explores all VACs used as they emerge, from A1 to C1 levels of proficiency. Using the COCA as a proxy for L1 usage, she also observes that the verbs associated with particular VACs move closer to L1 usage as proficiency increases. Aligning with previous studies, from both individual learners and bigger groups with the same L1 background (e.g. Römer and Garner 2019), Römer (2019) finds that lower level learners make use of a more restricted range of fixed verb associations which give way to a wider variety of associations at the higher levels of proficiency. The importance of both Römer and Garner (2019) and Römer (2019) is that they clearly underscore the need for a differentiated view of learner language. O'Keeffe (2021a) makes a case for the importance of building on Römer's findings to help guide and enhance DDL design using one of the patterns identified in the (2019) results. For our proposed research-informed framework, we also draw on the importance of learner corpus work on formulaicity, which we now examine, before presenting our three-stranded framework in sSection 4.

3.5. Formulaicity and learner language

The identification and description of formulaicity in language is seen as one of the overarching contributions of CL to the study of natural language (Forsberg Lundell, 2021). Erman and Warren (2000) estimate that over 58% of spoken and over 52% of written L1 English production is prefabricated in the mind of the user. From a learner perspective, this is one of the most challenging aspects of language learning, since, as Forsberg Lundell (2021: 371) notes, "formulaic language takes a long time to acquire".

Studies that look at the relationship between learner proficiency level and formulaic language use are growing. Forsberg and Bartning (2010) and Paquot (2018, 2019) found that formulaic language develops between B2 and C2 levels in L2 French and L2 English learners respectively and Forsberg Lundell (2021: 372) states that the results from these studies support the view that formulaic language is "a good indicator of second language proficiency especially at advance and very advanced levels". In UB developmental terms, in relation to the movement from formula to low-scope slot and frame to an abstracted system, formulaic language sits at the fully productive schematic end of the process. To get to the point of being able to subconsciously select, for example, *a huge amount of* over *a great amount of,* learners need to have experienced enough examples of usage "that their accidental and finite experience is truly representative of the total population of language of the speech community" that is, in terms of its "overall content, the relative frequencies of that content, and the mappings of form to functional interpretation." (Ellis, 2002: 167). Given the enormity of the L1 lexicon and breadth of possible constructions, it is therefore not surprising that L2 users might be distinguished by their ability or inability to use formulaic language in a fully productive way.

Collocation and colligation are core dimensions of formulaicity and CL research points to the need to focus on high frequency low cohesion collocations at lower levels (see Forsberg Lundell, 2021), while, at advanced levels, efforts need to go into mapping more register specific collocations that are not necessarily as high frequency but which are strong collocators (highly cohesive) (Granger and Bestgen 2014). Returning to Erman and Warren's (2000) point that over 58% of spoken and over 52% of written language is formulaic, it also points us to the need for learners to experience more spoken language patterns in DDL. In the next section, we present our framework for enhanced curation of patterns for DDL based on the aggregation of the strands of research that we have hitherto discussed in terms of its relevance to DDL (Sections 2 and 3.1–3.5).

3.6. Developing a framework for DDL

Distilling research from SLA and learner corpus studies with a view to abstracting what they offer for DDL, we propose the following framework of key principles (Fig. 2). These principles are based on findings on how patterns of language develop across levels of proficiency in terms of: acquisition, complexity and formulaicity. We argue that there are three key findings that need to become the basis for guiding principles for DDL as we discuss further below:

3.7. The acquisition principle: acquisition at lower levels differs to higher levels

Building on UB research discussed in Section 2, the Acquisition Principle is overarching: as learners move along this developmental pathway more new form-meaning mappings are acquired as their vocabulary grows and their awareness of phrasal combinations and co-selection evolves (Pérez-Paredes et al., 2020). This means lower-level learners need to work on refining knowledge about which words go together (and how these map to meanings). At the higher end of proficiency, there is a need to focus more on knowledge about how more than half of these combinations are usually formulaic and a need to gain breadth of repertoire in terms of fully abstracted (often figurative) patterns. An implication of this is that both tasks and data need to be mediated for levels, as we shall discuss in Section 4.

3.8. The complexity principle: there is movement from clausal development at lower levels to phrasal development at more advanced levels

As proficiency grows, learners move away from verb-based patterns and dependent clause complexity to phrasal complexity; the ability to use complex compressed noun phrases is a trait of development (see Section 3). Often, register awareness can be an important feature of this development where developing complexity in the noun phrase is a marker in second language acquisition from lower to higher level. Lower level learners are heavily reliant on topics (and tasks) when they put together sequences of words. As their language develops, this movement towards noun phrases co-occurs with an awareness and understanding

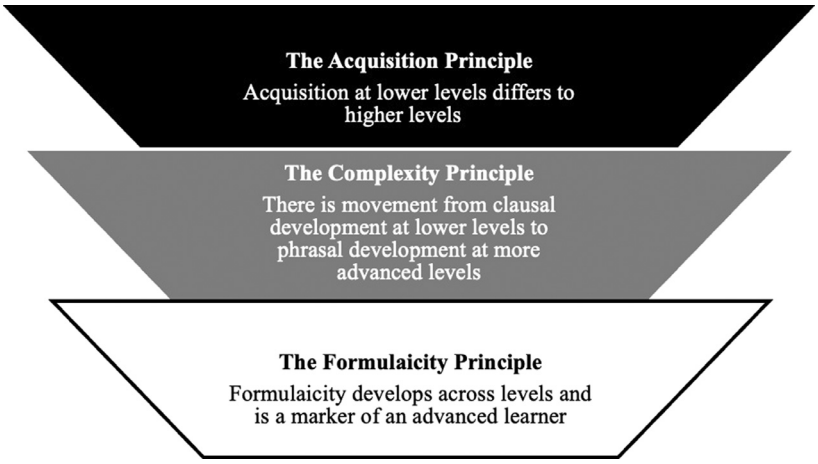


Fig. 2. Core principles for DDL design based on LCR and SLA theory.

Table 1
Top 10 Lemmatised collocates of *MAKE a/the* [] in Cambridge Learner Corpus A1 and C2 performance level data (using Sketch Engine) and BNC.

	CLC A1	Freq (PMW)	CLC C2	Freq (PMW)	BNC	Freq (PMW)
1	make a party	19 (0.32)	make a living	68 (1.12)	make a statement	382 (3.4)
2	make a pen-friend	13 (0.22)	make a lot	60 (1.0)	make a decision	245 (2.2)
3	make a lot	13 (0.22)	make the world	56 (1.0)	make a profit	203 (1.81)
4	make a birthday	12 (0.2)	make a difference	50 (0.83)	make a good	203 (1.81)
5	make a cake	8 (0.13)	make a decision	40 (0.67)	make a difference	174 (1.55)
6	make the food	5 (0.08)	make a choice	34 (0.56)	make a lot	152 (1.35)
7	make the concert	4 (0.07)	make a suggestion	24 (0.4)	make a living	144 (1.28)
8	make a concert	3 (0.05)	make the difference	20 (0.33)	make a note	123 (1.09)
9	make a dinner	3 (0.05)	make a person	19 (0.32)	make a point	123 (1.09)
10	make the project	2 (0.03)	make a career	17 (0.28)	make a contribution	116 (1.03)

of the discoursal function of the noun phrase and register. As we illustrate below (Section 4.2), DDL tasks for lower level learners can play a role in scaffolding development of noun phrase patterning and usage, especially with the affordances now available to work on discontinuous sequences. For higher levels, this principle points to the need to design tasks that augment the noun phrase repertoire, especially in terms of discourse function and register (e.g. noun phrases with an evaluative function in academic registers). This also relates to the Formulaicity Principle.

3.9. The formulaicity principle: formulaicity develops across levels and is a marker of an advanced learner

As discussed in Section 3, formulaicity is a pervasive feature of language and learners who do not appear to substantially acquire it until B2 level and beyond and, therefore, it is seen as a marker of advanced proficiency. Within a UB perspective, figuring out what words go together and mapping their meanings is often language specific and opaque in nature. In terms of differentiating for proficiency levels in the curation of patterns for DDL task design, this principle points to the importance of being aware that lower level learners will not have acquired many formulaic patterns while more advanced level learners need to accrue more formulaicity. As noted, collocation and colligation are also part of formulaic knowledge that needs to be fostered. This means at lower levels, working on collocation of very high frequency items (as well as high frequency formulaic patterns) (see Section 4.3). At higher levels, priority needs to be on highly cohesive low-frequency collocational patterns (often register-specific).

While we present three principles in our framework, in reality, they are interrelated. The first principle, the Acquisition Principle, overarches the Complexity and Formulaicity Principles. We now exemplify how the framework (and the findings on which they are based) can guide DDL design. This will include exemplars using a variety of interfaces, tools and data.

4. Applying key findings from SLA and LCR to DDL design: exemplars

4.1. Applying the acquisition principle: acquisition at lower levels differs to higher levels

If DDL is to be cognisant of UB findings in terms of the phases of acquisition, it means differentiating tasks and data by level:

- Pedagogical focus at A1 and A2 needs to be on fostering language experience so that learners enhance their knowledge of what words go together, i.e. the basic slots and frames.
- At B1 and B2 level, focus needs to shift more towards increasing slot and frame knowledge, i.e. enhanced both syntagmatic awareness of patterns and paradigmatic knowledge of what can go into certain slots in a pattern.
- As learners move towards C1 and C2 level, and have acquired a critical level competence in terms of their abstraction of patterns and meaning pairings, the focus needs to narrow to enhancing complexity in terms of collocational knowledge (especially in relation to lower frequency highly cohesive combinations), figurative meanings and phrasal complexity.

An important implication of the Acquisition Principle is the need to choose corpus data that suits the level of competency. To illustrate this point, Table 1 shows the patterns of the high-frequency verb *make* as a delexical verb in the British National Corpus (BNC) and compares it with the patterns that are evident in the A1 and C2 level CLC data.

If the A1 CLC results in Table 1 are a proxy for what A1 learners know about what words combine with *make*, then it shows us that:

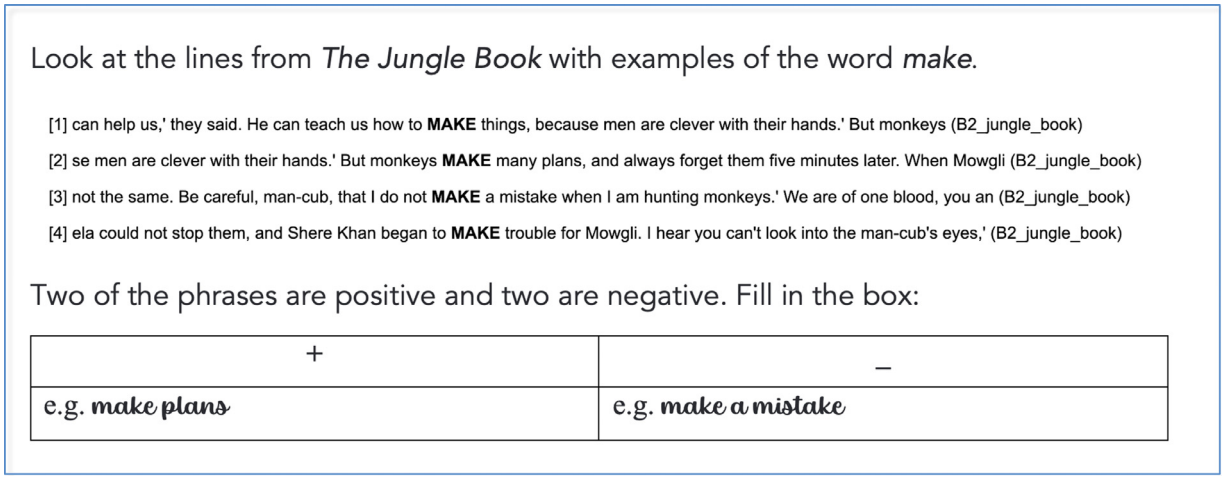


Fig. 3. Exemplar task with *make* using graded readers in *Lextutor*.

Civil and environmental engineering: This can **make a large difference in** the risk involved with each task if the proper tools are available to each worker.

Economics: However, the authors do not **make a claim for** greater applicability of their results, …

Economics: With a firm grasp of macroeconomic principles, one would be able to **make a more informed judgment of** the various economic arguments politicians put forth.

Fig. 4. Curated examples of collocates of *make* in MICUSP.

- They do not frequently use patterns with *make* and when they do, they have not quite worked out the correct patterns yet. In fact, A1 learners most commonly used patterns are often incorrect (e.g. *make a party, make a/the concert* = organise a concert).
- There is some evidence of formulaicity but it draws on the task rubric (13 uses of *make a pen-friend* (from task rubric).
- Literal patterns of *make* are found: *make* + *cake/food/dinner* etc.) to refer to real work situations and routines, and often these are incorrect (e.g. *make* + *party*).

In contrast, the C2 patterns with *make* evidence that:

- Learners frequently use a variety of patterns with *make.*
- Many of the BNC patterns are established and used (i.e. abstracted) in the C2 data (though not as frequently but this is exam data).
- C2 learners frequently use of figurative patterns (*make a living/difference/choice/suggestion/difference/career* etc.).

What Table 1 clearly shows is that an A1 learner would not benefit from free exploration of a native speaker corpus such as the BNC. What is required therefore is teacher mediation for the curation of data and tasks that will not overwhelm the learner and that focus on patterns that are differentiated to the level of the lower level learners.

Corpus tools and interfaces like *Antconc, SKELL, Voyant* allow user-friendly experiences if tasks are properly graded to level (see examples below). Other tools such as *Lextutor* offer access to graded texts. Fig. 3 shows a sample concordance task of *make* using *The Jungle Book* (using *Lextutor*) for A level learners.

Learner corpora also offer level-appropriate data sources (which can also align with development in register). For example, for intermediate (B1) level learners and upwards, guided search tasks can use register- and discipline-specific corpora (e.g. see Reppen and Olson 2020 who look at discipline-specific lexical bundles). The following examples of patterns with *make* in the Michigan Corpus of Upper Level Student Papers (MICUSP) illustrate this (Fig. 4).

Resources such as the EVP can also aid in differentiating tasks by level. Fig. 5 illustrates a small sample of patterns for C1/C2 level.

Taking just one of these items, for instance *make a name for yourself,* a C level learner can explore this figurative pattern further. Fig. 6 shows this pattern used with other pronouns in *The Movie Corpus* (https://www.english-corpora.org/movies/).

4.2. Applying the complexity principle: there is movement from clausal development at lower levels to phrasal development at more advanced levels

As noted, lower level writers rely more on verb-based patterns and dependent clause complexity while more proficient learners display more phrasal complexity, especially with increased use of complex compressed noun phrases. Noun phrase usage in lower levels relies on the topic to fill the 'noun slot'.

Even at lower levels, noun phrase complexity development is observed through an increase in range of determiners and adjectives premodifiers (e.g. typically in descriptions: *We have a big garden; There are so many people*.). The EGP (O'Keeffe and Mark 2017) can be used as a baseline description for noun usage across levels. Noun phrase competence at A1 and A2 levels are illustrated in Fig. 7.

As discussed, higher level writers use a wider repertoire of noun phrases in recurrent strings with an identifiable discourse function such as framing referential devices, used for time references, or evaluative or quantifying purposes with a following noun phrase, e.g. *the end of the* + noun phrase*, the number of* + noun phrase*, a large number of* + noun phrase*, a great opportunity for* + noun phrase*, a good deal of* + noun phrase. Mindful of Gray and Biber's (2103) call to examine discontinuous sequences there is a need to draw lower proficiency learners' attention to a wider variety of uses and more complex patterning of the noun phrase.

AntConc 4.0 software (Anthony, 2022) now allows us to search for discontinuous sequences. This is an important development. Fig. 8

Base Word	Guideword	Level	Part of Speech	Topic	Details
make a nonsense of sth		C2	phrase		Details
make a name for yourself		C2	phrase		Details
make a point of doing sth		C1	phrase		Details
make a note of sth		C1	phrase		Details
make a go of sth		C2	phrase		Details

Fig. 5. Sample of items listed in EVP for C level learners under *make*.

3	2016	US/CA	Nerdland	and I are going to do something special. We're going to make a name for ourselves. Will you go away now? Anyhow, n-nice seeing you.
4	2016	US/CA	Nerdland	John: Welcome to cyber century, friend. We're going to make a name for ourselves without even setting foot into the real world. How are we going
5	2015	US/CA	Red Herring	was trying to make another play for Vegas, they'd want to make a name for themselves, right? Absolutely. Well what better way to do that than
6	2015	US/CA	In Football We Trust	lot of pressure. Leva: He's done. Just trying to make a name for myself. On the field, man, I ain't that nice.
7	2015	US/CA	In Football We Trust	bunch of kids trying to, you know, just basically trying to make a name for themselves. I don't mean to be negative, but the Bloomfield name
8	2015	US/CA	Forsaken	Men from far and wide are gon na come looking for me to make a name for themselves. Don't go away. I can't finish the field on
9	2015	US/CA	Kill or Be Killed	No, it's not. I spent all these years trying to make a name for myself, and then they go make a mockery out of me like that
10	2015	US/CA	Dead Rising: Watchtower	come here to explore people. This is how we are going to make a name for ourselves. You know what? Here. What's this? Remind yourself
11	2015	US/CA	For Grace	Today or tomorrow? Tomorrow. A young chef who's trying to make a name for themselves, Are always going to be cast upon the shadow Of someone wh
12	2015	US/CA	Sugar Babies	you call no one. Do you understand me? You want to make a name for yourself in this business, don't you? You don't want to
13	2015	US/CA	Reel Rock 10	from a small town in Maryland and I wanted to recreate myself. Make a name for myself. Do things that people were impressed by. (SCREAM) Emerson
14	2015	US/CA	Merry Matrimony	After school, I thought if I could get one big job. Make a name for myself, I could come back and sweep Brie off her feet. We
15	2014	US/CA	Sharktopus vs. Ptera...	I knew, this is my chance to do real science, to make a name for myself. But it's just so unpredictable. You mean dangerous. I
16	2013	US/CA	The Wolf of Wall Street	the way we do things. You got ta understand, trying to make a name for ourselves. But, I want you to understand, we don't do

Fig. 6. Extract from concordances of *make a name for* using English-Corpora.org *The Movie Corpus*.

SuperCategory	SubCategory	Level	Can-do statement	Example	Details
NOUNS	noun phrases	A1	FORM: DETERMINER + NOUN Can form simple noun phrases with a limited range of determiners + singular and plural nouns. ▶ Nouns and noun phrases: functions	Example	Details
NOUNS	noun phrases	A1	FORM: DETERMINER + ADJECTIVE + NOUN Can form simple noun phrases by pre-modifying singular and plural nouns with an adjective after a determiner.	Example	Details
NOUNS	noun phrases	A1	FORM: ADJECTIVE + PLURAL NOUN Can form simple noun phrases by pre-modifying plural nouns with an adjective and no determiner.	Example	Details
NOUNS	noun phrases	A1	FORM: NOUN + NOUN Can form noun phrases by pre-modifying a limited range of nouns with another noun.	Example	Details
NOUNS	noun phrases	A2	FORM: DETERMINER + UNCOUNTABLE NOUN Can form simple noun phrases with a limited range of determiners + uncountable nouns.	Example	Details
NOUNS	noun phrases	A2	FORM: DETERMINER + NOUN Can form simple noun phrases by pre-modifying nouns with an increasing range of determiners.	Example	Details
NOUNS	noun phrases	A2	FORM: NOUN PHRASES WITH ADJECTIVES Can pre-modify noun phrases with a limited range of more than one adjective.	Example	Details

Fig. 7. Screenshot of a sample of A-level learners' noun phrase descriptors from the EGP.

shows results of 3-gram search with one open slot using a pre-loaded corpus within *Antconc 4.0*.

In line with Gray and Biber (2013), we also see that these discontinuous sequences consist mostly of function words. Our example concordance task (Fig. 9) takes the most frequent sequence *the* + *of* and using the *AntConc* preloaded corpus uses *the end of* as an example for use of DDL with A level learners, also drawing on the EVP to filter the A level use of *end*.

As a result of the task, learners might draw the following conclusions:

- *End* in these examples means 'the final part of something such as a period of time'.
- *End* with this meaning is part of a sequence ***the end of the.***
- We use ***the end of the*** for
  - A general time period: the end of the **century/year/month/day.**
- A part of the day: the end of the **night/morning.**
  - A season: the end of the **summer.**
  - An event: the end of the **war.**
- We often use ***at*** *the end of the* and we sometimes use ***by*** *the end of the*

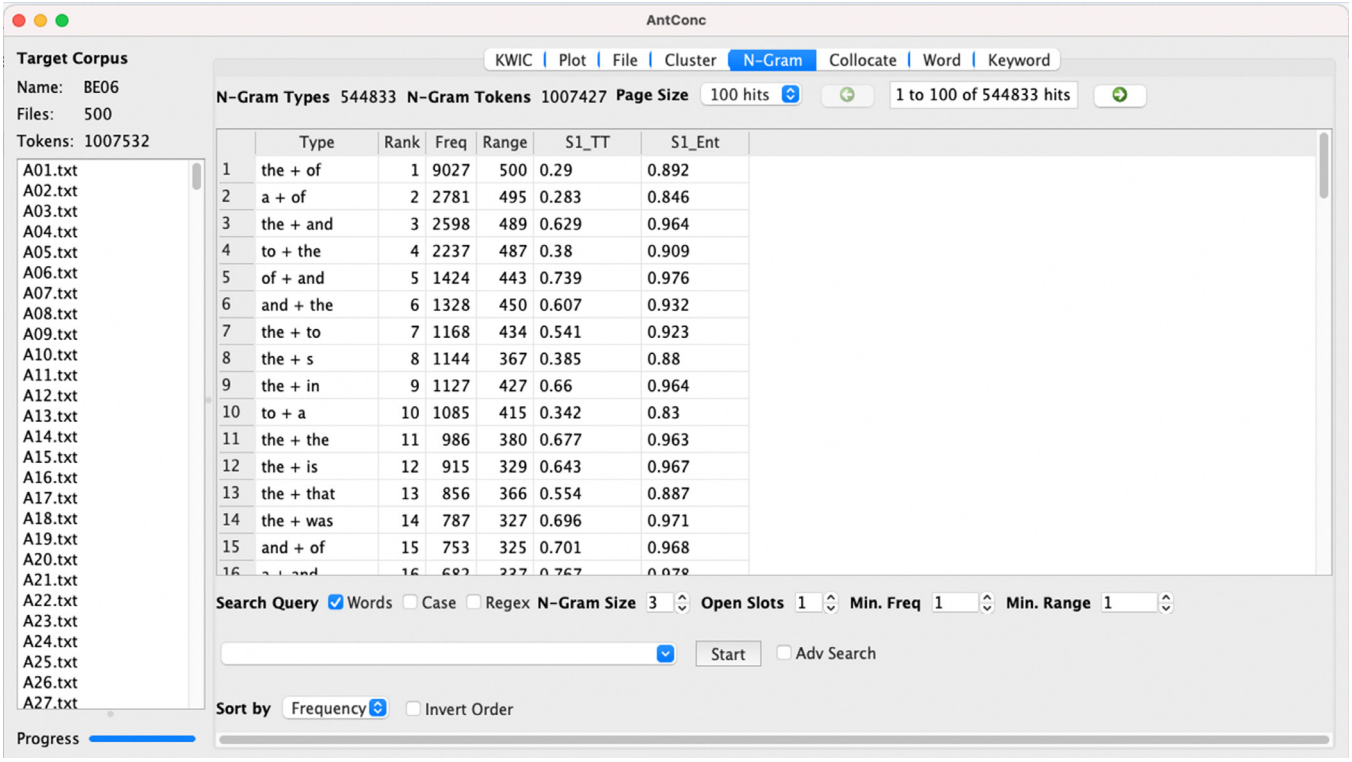


Fig. 8. Top 15 results of 3-grams including one open slot from *AntConc 4.0* in pre-loaded corpus.

Antconc 4.0 provides a way to explore discontinuous patterns to see which words most frequently 'fill' slots and which words most frequently occur around these words. This kind of guided exploration is accessible to learners of all levels and will help in building up awareness of phrasal complexity even at lower levels. Curated learner corpus examples from the level of the learners might offer a means of ensuring the examples are differentiated appropriately.

4.3. Applying the formulaicity principle: formulaicity develops across levels and is a marker of an advanced learner

As noted above, learners at an advanced level will have gained high frequency low cohesion patterns but need to develop more high cohesion high *and* low frequency patterns. There is therefore scope to build advanced learners' repertoires so that they work on high frequency items to build up more formulaic patterning. The COCA interface allows for a gradation of collocates into 'loose', 'medium' and 'tight' (see Fig. 13) and this can be exploited to develop patterns across different levels of cohesion. Fig. 10 illustrates and example of a task designed to work on these patterns.

BFy way of commentary of how this task might work using the verb *come:*

Part 1: Fig. 11 gives a small sample from the EVP filtered for C1 / C2 level uses. Each is hyperlinked to corpus-based examples, offering learners many options to follow up on.

Part 2: Looking at the collocates of *come* will bring advanced learners' focus to the importance of collocations (Fig. 12) such as *come to the **conclusion**; come as a **surprise**; come in **handy**; come to **mind**,* etc. Importantly, it will give condensed exposure to high frequency formulaic items and their figurative uses.

In the follow-up activity (2c), use is made of the English-corpora.org 'clusters' tab (Fig. 13) which generates lists based on collocation and colligational strength (referred to as *loose, medium* and *tight*). Advanced learners can push their repertoire by working on *tight* clusters (i.e. high association and usually lower frequency).

Free foraging by learners can lead them down interesting paths. For example, if they click on *come as a surprise* in the 'tight clusters' results for *come,* they can explore the functions of this phrase in context and possibly see the negative prosody that pertains to it (Fig. 14). This could further be compared with *came as no surprise,* and so on.

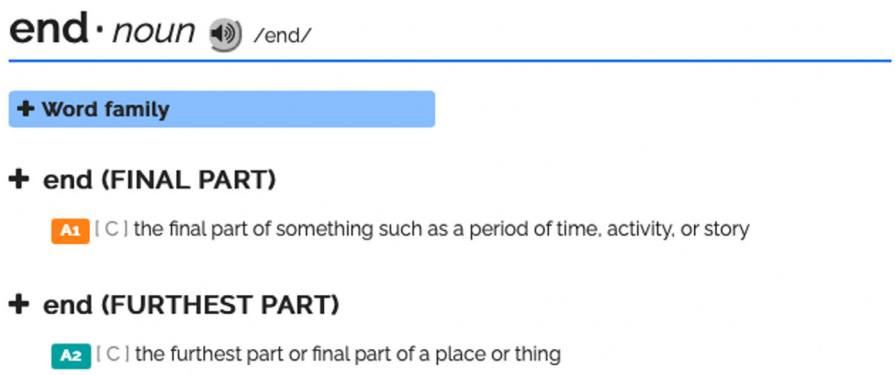
5. Conclusion

Based on an aggregation of key findings from SLA and LCR, we have proposed a framework to enhance DDL curation and task design. This framework moves away from the original ethos of DDL in which the student engaged in a discovery process of inductive learning. While this is still an attainable ideal, we argue that if DDL is to work across levels, there is need for principled mediation to differentiate data and tasks for different levels. More conceptualisation about the nature of teaching and learning in DDL is required as part of this mediation process (O'Keeffe, 2021b attempts to do this). The nuances of teacher and peer mediation in DDL tasks have a direct relationship with the degree of freedom or "free-range-ness" (Fig. 15).

Informed by UB research on language development, our framework for DDL can offer a principled basis for differentiated tasks and mediated data so as to move focus from basic formula (word combinations) to slot and frame sequences to fully abstracted constructions. As learners move up levels, more focus can be put on acquiring new meanings and narrowing in on phrasal word association, co-selection, collocation and formulaicity.

As Table 2 illustrates, the three principles in our framework, based on acquisition, complexity and formulaicity take cognisance of the stages and process of development and acquisition so as to guide the tailoring and mediation of tasks and data. Therefore at lower levels of proficiency, there is a need to curate patterns as input which align with high frequency items in a corpus (Table 2). As learners develop and acquire word combinations and slot and frame patterns, from which

1a Look at these definitions of the word 'end' from English Vocabulary Profile.



1b Now look at these examples with *end*. Which definition matches their uses?

ls of output. 'This isn't	the end of	the story. It
vitable consequence of	the end of	the Cold War."
ive his decision before	the end of	the year. Crack
missions to the RAE by	the end of	the month. Panels
30 against the euro by	the end of	the year, in
viser to the board after	the end of	the year. Revenues
w album X, released at	the end of	the month. Her
e disco and dancing at	the end of	the night. Zebra
imultaneous attacks at	the end of	the morning rush

2 What do you notice about the patterns of words around *end* ?

Fig. 9. Exemplar task with patterns around *end* using the EVP and *AntConc 4.0* in pre-loaded corpus.

Table 2
Acquisition - Complexity - Formulaicity Framework for DDL design.

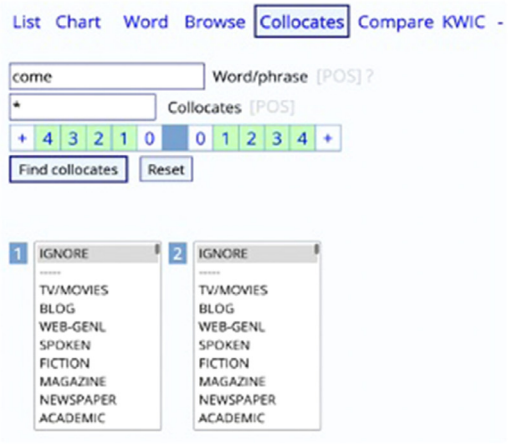
Principle	A1 A2	B1 B2	C1 C2
	⟶	⟶	⟶
Acquisition	word combinations	identifying slots and frames	formulaic knowledge: abstracted patterns narrowing of word associations register awareness
Complexity	focus on verb-based patterns	moving from verb- to noun-based patterns	focus on noun complexity
Formulaicity	focus on literal meaning	broadening from literal meaning to figurative meaning	focus on figurative meaning

1 Here are the top 20 most frequent verbs in the British National Corpus. Work in pairs and choose one verb to explore using *English Vocabulary Profile* (EVP) https://www.englishprofile.org/ and English-Corpora.org:

do	take	find	tell
see	know	think	let
get	say	look	use
go	give	help	keep
make	come	put	like

- Look up the verb in the EVP and filter for C1 and C2 level. List five new ways of using the verb that you have found in EVP.
- Examine these five new phrases using a corpus and prepare a mini-presentation for your class.

2 Look up one of the verbs in the COCA using the **Collocates** function with the following settings. Then find the answers to the questions a, b and c.



a What is the most frequent collocate that you found and what does the pattern mean?
b Note five new patterns and meanings that you have identified from the lists of collocates.
c Within the **Collocates** results page, click on the **Clusters** tab on the top right corner. Fill out the following grid based on the *loose*, *medium* and *tight* clusters of the verb you are exploring (add an example that helps you remember its meaning):

Top 3 loose cluster patterns	Top 3 medium cluster patterns	Top 3 tight cluster patterns

Fig. 10. Exemplar task for finding cluster patterns (by cohesion) using COCA.

come about		C2	phrasal verb	
come between sb		C2	phrasal verb	relationships
come across	SEEM	C1	phrasal verb	people: personality
come into sth		C2	phrasal verb	money
come to think of it		C2	phrase	communication
come out	BE SAID	C2	phrasal verb	communication
come round	BECOME CONSCIOUS	C1	phrasal verb	body and health
how come		C1	phrase	communication
come first		C1	phrase	
come in handy		C2	phrase	
come under fire	IDIOM	C2	phrase	

Fig. 11. Sample of some of the EVP entries for patterns with high frequency verb *come* at C level.

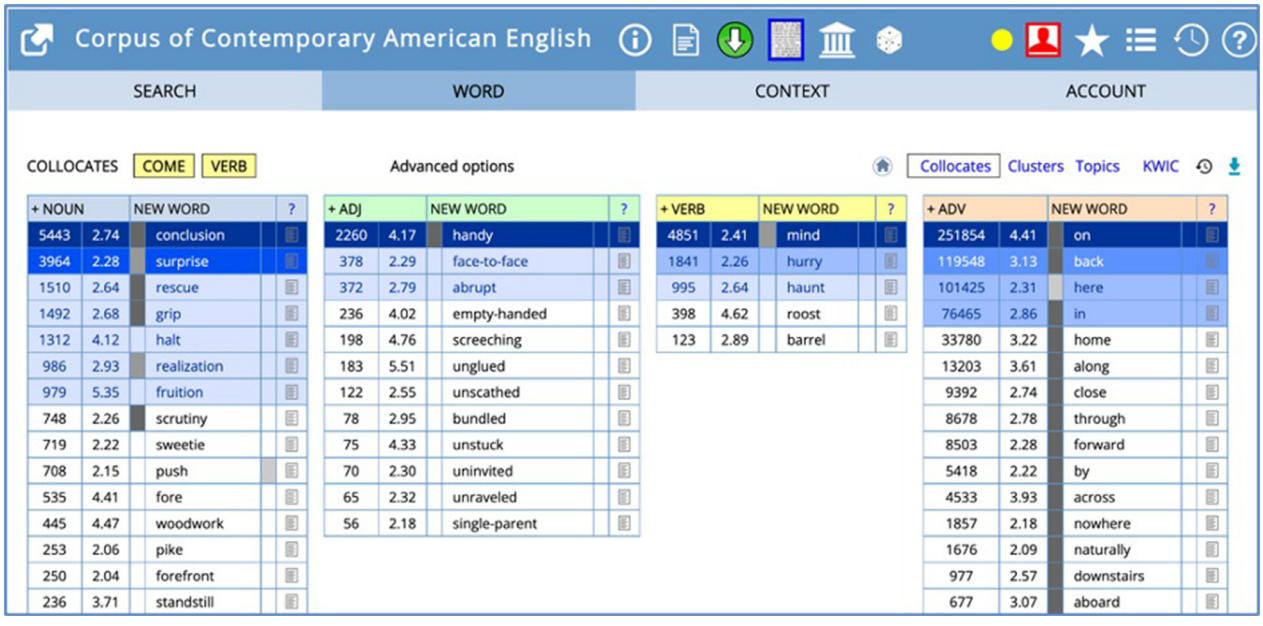


Fig. 12. Collocation patterns of *come* using English-corpora.org collocates function.

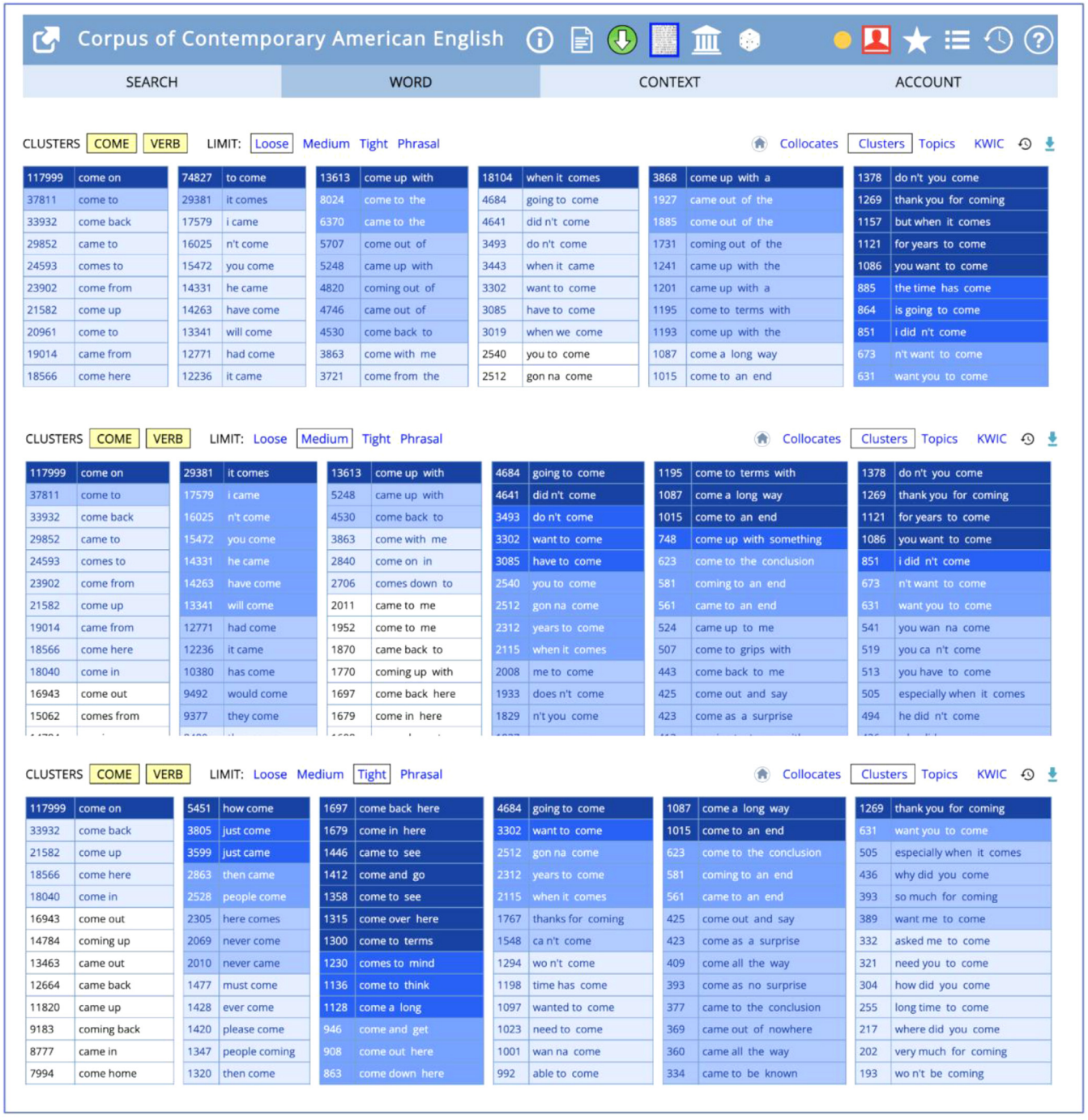


Fig. 13. Sample of partial screenshots for *loose, medium* and *tight* clusters with *come* using *English-corpora.org*.

they eventually abstract more new meanings, development narrows to phrasal word association and co-selection. In other words, learners refine knowledge about which words go together and how these map in their meanings. They eventually acquire more and more knowledge of formulaic patterns (Table 2).

This is by no means a fully-formed framework but, as a starting point, it can be expanded. Any one of the three overarching principles could be given a more fine-grained treatment. In relation to the Acquisition Principle, the following need further attention: the acquisition stage of low proficiency learners and the selection of verb patterns based on frequency and meaning; the sequencing of VAC patterns in terms of slot and frame development for B level learners; the noun complementation patterns that need to be prioritised for C level learners, etc. Other important steps in evolving our proposed framework need to also investigate the importance of register as a mediating factor in the Complexity and Formulaicity Principles. Research underpinning the Complexity Principle for example is based on findings from written learner data, leading to the assumption that language production will become more phrasal with higher proficiency levels, but this need to be tested on spoken discourse (particularly spoken discourse that is not informationally driven). Additionally, in relation to the Formulaicity Principle, the nuances of what formulaicity looks like at advanced levels of speech (e.g. more fixed expressions) vs. writing (more variable slots) needs further exploration using spoken learner corpora.

Microsoft) have all introduced Metro to the larger masses the new logo does not come as a surprise. It falls perfectly in place with what we've been seeing
she is famous! HOP OFF HER NUTS!! # This shouldn't really come as a surprise to anyone. And this became everyone's business when they made
's changed his stance on so many issues, being called a liar shouldn't come as a surprise to anyone at all. Romney will say whatever it takes,
a lunch break initiative program that proved to be a great success. What might come as a surprise to you as it did me: the logistics of making sure
points is critical. Let's view this in greater detail. # This may come as a surprise, but one of my secondary career occupations has been working for
sharp guy. # KSTP's reaction? # KTLK's launch " does not come as a surprise to us, " said KSTP general manager Todd Fisher in a
on gut-shot: # The launch of the new FM talk format also didn't come as a surprise to Carol Grothem, broadcast manager for Campbell Mithun ad agency.
people who work for non-profit organisations never commit violent crimes is laughable. This may come as a surprise, but criminals come from all walks of
went along these lines too, but its article added that this endorsement did not come as a surprise and that for this reason it is not that much of a
3158612 # It shouldn't have come as a surprise to learn that after getting into a major car accident during the
I understand that. I have been one of those writers. This may not come as a surprise, at this point in this essay, but for a long
and her desire to return to kickboxing and Muay Thai. So it does not come as a surprise to learn that she will be defending one of her Muay Thai
courses have site projects and require residency so let this pharmacy around the world not come as a surprise to you. # Most online colleges offer financi
has meant that disposal of carcasses takes much longer than before. # It might come as a surprise to your readers that the most famous Dakhma in India
behind. Now with its manager banned for the next 10 months it would not come as a surprise if Juventus was crowned champions next spring, taking hom

Fig. 14. Sample of concordance lines for *came as a surprise* in the Corpus of Contemporary English using *English-corpora.org*.

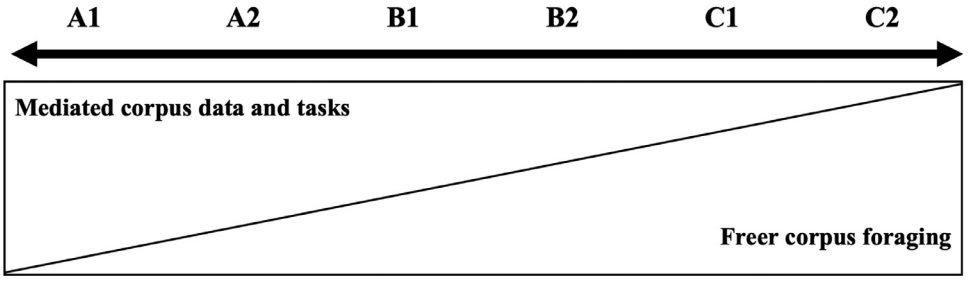


Fig. 15. The cline from mediation of data and task to free foraging in DDL.

The holy grail would be to arrive at a list of 'pathbreaking' patterns (after Ninio, 1999, see Section 2) for DDL. However, this would need to be backed up by SLA experimentation. Existing work on pattern grammar (Hunston and Francis 2000), VACs (e.g. Ellis et al. 2014), lexical bundles (e.g. Biber and Gray, 2016), grammar patterns and semantic frames (Perek and Patten 2019), as well as corpus-based resources (such as the English Grammar Profile, the English Vocabulary Profile, *AntConc* and *Lextutor*), can all feed into this. Advances from here hinge on closer alliance between Second Language Acquisition, Learner Corpus Research and Data-Driven Learning.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Alexopoulou, T., Geertzen, J., Korhonen, A., Meurers, D, 2015. Exploring big educational learner corpora for SLA research: perspectives on relative clauses. Int. J. Learn. Corpus Res. 1 (1), 96–129.
Anthony, L., 2022. AntConc (Version 4.0.10) [Computer Software]. Waseda University, Tokyo, Japan Available from.
Biber, D., Gray, B., 2011. Grammatical change in the noun phrase: the influence of written language use. English Lang. Linguist. 15 (2), 223–250.
Biber, D., Gray, B., 2016. Grammatical Complexity in Academic English: Linguistic Change in Writing. Cambridge University Press.
Biber, D., Conrad, S., Cortes, V., 2004. *If you look at…*: lexical bundles in university teaching and textbooks. Appl. Linguist. 25, 371–405.
Biber, D., Gray, B, Poonpon, K., 2011. 'Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?'. TESOL Q. 45 (1), 5–35.
Biber, D., Reppen, R., Staples, S., Egbert, J., 2020. Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. Int. J. Learn. Corpus Res. 6, 38–71.
Boulton, A., 2017. Corpora in language teaching and learning. Lang. Teach. 50 (4), 483–506. doi:10.1017/S0261444817000167.
Boulton, A., Cobb, T, 2017. Corpus use in language learning: a meta-analysis. Lang. Learn. 67 (2), 348–393.
Capel, A., 2010. A1–B2 vocabulary: insights and issues arising from the English profile wordlists project. Engl. Profile J. 1, 2–6.
Chen, Y., Baker, P., 2010. 'Lexical bundle in L1 and L2 academic writing'. Lang. Learn. Technol. 14/2, 30–49.
Cobb, T., Boulton, A., 2015. Classroom applications of corpus analysis. In: Biber, D., Reppen, R. (Eds.), Cambridge Handbook of Corpus Linguistics. Cambridge University Press, Cambridge, pp. 478–497.
De Smet, H., Cuyckens, H., 2007. Diachronic aspects of complementation: constructions, entrenchment and the matching-problem. In: Cain, C., Russom, G. (Eds.), Studies in the History of the English Language III: managing chaos: Strategies for identifying change in English. De Gruyer Mouton, Boston/Berlin, pp. 1–37.
Díez-Bedmar, M.B., Pérez-Paredes, P., 2020. Noun phrase complexity in young Spanish EFL learners' writing: complementing syntactic complexity indices with corpus-driven analyses. Int. J. Corpus Linguistics 25 (1), 4–35.
Dulay, H.C., Burt, M.K., 1973. Should we teach children syntax? Lang. Learn. 23 (2), 245–258.
Ellis, N.C., 2002. Frequency effects in language processing a review with implications for theories of implicit and explicit language acquisition. Stud. Sec. Lang. Acquis. 24, 143–188.
Ellis, N.C., 2003. Constructions, chunking, and connectionism: the emergence of second language structure. In: Doughty, C., Long, M.H. (Eds.), Handbook of Second Language Acquisition. Blackwell, Oxford, pp. 33–68.
Ellis, N.C., 2012. Frequency effects. In: Robinson, P. (Ed.), The Routledge Encyclopedia of Second Language Acquisition. Routledge, New York, pp. 260–265.
Ellis, N.C., Larsen-Freeman, D, 2006. Language emergence: implications for applied linguistics. Introduction to the special issue. Appl. Linguist. 27 (4), 558–589.
Ellis, N.C., Römer, U., O'Donnell, M.B., 2016. Constructions and usage-based approaches to language acquisition. Lang. Learn. 66, 23–44.
Erman, B., Warren, B, 2000. 'The idiom principle and the open choice principle' Text-Interdisciplinary. J. Study Discourse 20 (1), 29–62.
Flowerdew, L., 2015. Data-driven learning and language learning theories: whither the twain shall meet. In: Leńko-Szymańska, A., Boulton, A. (Eds.), Multiple Affordances of Language Corpora for Data-Driven Learning. John Benjamins, Amsterdam, Netherlands, pp. 15–36.

Forsberg Lundell, F., 2021. Formulaicity and corpora. In: Tracy-Ventura, N., Paquot, M. (Eds.), Routledge handbook of second language acquisition and corpora. Routledge, London, pp. 370–381.
Forsberg, F., Bartnin, I., 2010. 'Can linguistic features discriminate between the communicative CEFR-levels?: a pilot study of written L2 French'. In: Bartning, I., Martin, M., Vedder, I. (Eds.), Communicative proficiency and linguistic development, pp. 133–158 Eurosla monograph series 1.
Gablasova, D., Brezina, V., McEnery, T., 2019. The trinity lancaster corpus: development, description and application. Int. J. Learn. Corpus Res. 5 (2), 126–158.
Gabrielatos, C., 2005. Corpora and language teaching: Just a fling or wedding bells? Teach. Engl. Sec. Foreign Lang. 8 (4), 1–34.
Garner, J., 2016. A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. Int. J. Learn. Corpus Res. 2 (1), 31–67.
Goldberg, A.E., 2006. Constructions at Work. The Nature of Generalization in Language. Oxford University Press, Oxford.
Granger, S., Bestgen, Y., 2014. The use of collocations by intermediate vs. advanced non–native writers: a bigram-based study. Int. Rev. Appl. Linguist. Lang. Teach. 52 (3), 229–252.
Gray, B., Biber, D, 2013. 'Lexical frames in academic prose and conversation'. Int. J. Corpus Linguist. 18 (4), 109–136.
Green, A., 2010. Requirements for reference level descriptions for English. Engl. Profile J. 1.
Harrison, J., Barker, F. (Eds.), 2015. *English Profile in Practice. English Profile Studies, Vol. 5*. Cambridge University Press, Cambridge.
Hawkins, J.A., Buttery, P., 2010. 'Criterial features in learner corpora: theory and illustrations'. Engl. Profile J. 1 (1), 1–23.
Hawkins, J., Filipović, L., 2012. Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework. Cambridge University Press, Cambridge.
Johansson, S., 2009. 'Some thoughts on corpora and second-language acquisition'. In: Aijmer, K. (Ed.), Corpora and Language Teaching. John Benjamins, Amsterdam, pp. 33–44.
Langacker, R.W., 1987. Foundations of Cognitive Grammar: Theoretical Prerequisites, 1. Stanford University Press.
Lee, H., Warschauer, M., Lee, J.H., 2019. The effects of corpus use on second language vocabulary learning: a multilevel meta-analysis. Appl. Linguist. 40 (5), 721–753.
Lenko-Szymanska, A., 2014. The acquisition of formulaic language by EFL learners: a cross-sectional and cross-linguistic perspective. Int. J. Corpus Linguist. 19 (2), 225–251.
Meunier, F., 2015. 'Second language acquisition theory and learner corpus research'. In: Granger, S., Gilquin, G., Meunier, F. (Eds.), The Cambridge Handbook of Learner Corpus Research (2015). Cambridge University Press.
Meunier, F., 2020. A case for constructive alignment in DDL: Rethinking outcomes, practices and assessment in (data-driven) language learning. In: Crosthwaite, P. (Ed.), Data-driven Learning for the Next Generation. Corpora and DDL for Pre-Tertiary Learners. Routledge, pp. 13–30.
Murakami, A., Alexopoulou, T, 2016. L1 influence on the acquisition order of English grammatical morphemes. Stud. Sec. Lang.. Acquisit. 38 (03), 365–401.
Ninio, A., 1999. Pathbreaking verbs in syntactic development and the question of prototypical transitivity. J. Child Lang. 26 (3), 619–653.
Ninio, A., 2005. Testing the role of semantic similarity in syntactic development. J. Child Lang. 32 (1), 35–61.
O'Keeffe, A., 2021a. 'Data-driven learning and the second language acquisition interface debate'. In: Peréz-Paredes, P., Mark, G. (Eds.), Beyond the Concordance: Multiple Applications of Language Corpora for Language Education. John Benjamins, Amsterdam, pp. 35–55.
O'Keeffe, A., 2021b. Data-driven learning – a call for a broader research gaze. Lang. Teach. 54 (2), 259–272.
O'Keeffe, A., Mark, G., 2017. The English grammar profile of learner competence: methodology and key findings. Int.l J. Corpus Linguist. 22 (4), 457–489.
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.
Paquot, M., 2018. Phraseological competence: a useful toolbox to delimitate CEFR levels in higher education? Insights from a study of EFL learners' use of statistical collocations. Lang. Assessm. Q. 15, 29.
Paquot, M., 2019. The phraseological dimension in interlanguage complexity research. Sec. Lang. Res. 35 (1), 121–145.
Pérez-Paredes, P., 2010. 'Corpus linguistics and language education in perspective: appropriation and the possibilities scenario. In: Harris, T., Moreno Jaén, M. (Eds.), Corpus Linguistics in Language Teaching. Peter Lang, Frankfurt, pp. 53–73.
Pérez-Paredes, P., Díez-Bedmar, M.B., 2019. Certainty adverbs in spoken learner language. The role of tasks and proficiency. Int. J. Learn. Corpus Res. 5 (2), 252–278.
Pérez-Paredes, P., Mark, G., O'Keeffe, A., 2020. *The Impact of Usage-Based Approaches on Second Language Learning and Teaching.* Cambridge Education Research Reports. Cambridge University Press, Cambridge.
Reppen, R., Olson, S., 2020. Lexical bundles across disciplines: a look at consistency and variability. In: Römer, U., Cortes, V., Friginal, E. (Eds.), Advances in Corpus-based Research on Academic Writing: Effects Of Discipline, Register and Writer Expertise. John Benjamins, Amsterdam, pp. 169–182. doi:10.1075/scl.95.07rep.
Römer, U., 2019. A corpus perspective on the development of verb constructions in second language learners. Int. J. Corpus Linguist. 24 (3), 268–290.
Römer, U., Garner, J.R., 2019. The development of verb constructions in spoken learner English: tracing effects of usage and proficiency. Int. J. Learn. Corpus Res. 5 (2), 206–229.
Römer, U., O'Donnell, M.B., Ellis, N.C., 2014. Second language learner knowledge of verb–argument constructions: effects of language transfer and typology. Mod. Lang. J. 98 (4), 952–975.
Römer, U., Skalicky, S., Ellis, N.C., 2018. Verb-argument constructions in advanced L2 English learner production: insights from corpora and verbal fluency tasks. Corpus Linguist. Linguist. Theory.
Sharwood Smith, M., 1993. Input enhancement in instructed SLA. Stud. Sec. Lang. Acquisit. 15 (2), 165–179.
Staples, S., Egbert, J., Biber, D., McClair, A., 2013. 'Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section'. J. Engl. Acad. Purp. 12/3, 214–225.
Thewissen, J., 2013. Capturing L2 accuracy developmental patterns: insights from an error-tagged EFL learner corpus. Mod. Lang. J. 97, 77–101.
Tono, Y., Díez-Bedmar, M.B., 2014. Focus on learner writing at the beginning and intermediate stages: the ICCI corpus. Int.l J. Corpus Linguist. 19 (2), 163–177.
Vidakovic, I., Barker, F., 2010. Use of words and multi-word units in Skills for Life writing examinations. Univ. Camb. ESOL Exam. Res. Notes (41) 7–14.
Vyatkina, N., Boulton, A. (Eds.), 2017. Corpora in language teaching and learning. special issue. Lang. Learn. Technol. 21 (3).
Wulff, S., Ellis, N.C., 2018. In: Usage-based Approaches to Second Language Acquisition, 54. John Benjamins, Amsterdam, pp. 37–56.
Zipf, G.K., 1935. The Psycho-Biology of Language. Houghton, Mifflin, Oxford, England.