16

Learner corpora and formulaic language in second language acquisition research

Nick C. Ellis, Rita Simpson-Vlach, Ute Römer,
Matthew Brook O'Donnell and Stefanie Wulff

1 Introduction

Just how proficient are second language learners in using formulaic language? Do formulaic phrases play a role in second language acquisition (SLA)? These are the two questions to be addressed here using evidence from learner corpus research. Whilst Krashen and Scarcella (1978) argued that formulaic language was outside the creative language process, Ellis (1996) proposed that learners' long-term knowledge of lexical sequences in formulaic phrases serves as the database for language acquisition. The current chapter addresses the apparent paradox whereby analyses of learner language show that second/foreign (L2) learners typically do not achieve native-like formulaicity and idiomaticity (Pawley and Syder 1983; Granger 1998b), whereas longitudinal analyses of learner corpora such as Myles (2004) show that formulaic phrases can provide learners with complex structures beyond their current grammar, and that resolving the tension between these grammatically advanced chunks and the current grammar drives the learning process forward.

Usage-based theories of language hold that L2 learners acquire constructions from the abstraction of patterns of form–meaning correspondence in their usage experience and that the acquisition of linguistic constructions can be understood in terms of the cognitive science of concept formation following the general associative principles of the induction of categories from experience of the features of their exemplars

(Robinson and Ellis 2008; Hoffmann and Trousdale 2013). In natural language, the type–token frequency distributions of the occupants of each part of a construction, their prototypicality and generality of function in these roles, and the reliability of mappings between these, all affect the learning process. Child-language researchers (Tomasello 2003; Lieven and Tomasello 2008) and L2 researchers (Ellis 2013) have proposed that formulaic phrases with routine functional purposes play a large part in this experience, and the analysis of their components gives rise to abstract linguistic structure and creativity: '[t]he typical route of emergence of constructions is from formula, through low-scope pattern, to construction' (Ellis 2002: 143).

Researching these issues necessitates the bringing together of a range of types of methods to triangulate with learner corpus research (see also Chapter 3, this volume). Learner corpora are essential in showing the evidence of learner formulaic use, and dense longitudinal corpora allow the charting of the growth of learner use (Paquot and Granger 2012). But the analysis of large corpora of everyday usage like the *British National Corpus* (*BNC*)[1] and the *Corpus of Contemporary American English* (*COCA*)[2] is a necessary adjunct in order to get a picture of typical language experience which serves learners as their evidence for learning (McEnery and Hardie 2012). Furthermore, psycholinguistic experiments are necessary to look at learners' implicit knowledge of linguistic structures and the strengths of association of their components as they affect on-line processing in language comprehension and production (e.g. Ellis 2002; Schmitt 2004; see also Chapter 4, this volume). We concur with Gilquin and Gries (2009: 9) that '[b]ecause the advantages and disadvantages of corpora and experiments are largely complementary, using the two methodologies in conjunction with each other often makes it possible to (i) solve problems that would be encountered if one employed one type of data only and (ii) approach phenomena from a multiplicity of perspectives'.

2 Core issues

2.1 L2 processing is sensitive to the statistical properties of formulaic language

Research in psycholinguistics, corpus linguistics and cognitive linguistics demonstrates that language users have rich knowledge of the frequencies of forms and of their sequential dependencies in their native language (Ellis 2002). Language processing is sensitive to the sequential probabilities of linguistic elements at all levels from phonemes to phrases, in comprehension as well as in fluency and idiomaticity of speech production.

[1] www.natcorp.ox.ac.uk/ (last accessed on 13 April 2015).

[2] http://corpus.byu.edu/coca/ (last accessed on 13 April 2015).

This sensitivity to sequence information in language processing is evidence of learners' implicit knowledge of memorised sequences of language, and this knowledge serves as the basis for linguistic systematicity and creativity. The last ten years have seen substantial further research confirming native and L2 users' implicit knowledge of linguistic constructions and their probabilities of usage (Ellis 2012a; Rebuschat and Williams 2012). Illustrative recent studies demonstrating second language learners' implicit knowledge of the sequential probabilities of linguistic elements include the following.

Jiang and Nekrasova (2007) examined the representation and processing of formulaic sequences using on-line grammaticality judgement tasks. English as a second language speakers and native English speakers were tested with formulaic and non-formulaic phrases matched for word length and frequency (e.g. *to tell the truth* vs *to tell the price*). Both native and non-native speakers responded to the formulaic sequences significantly faster and with fewer errors than they did to non-formulaic sequences.

Conklin and Schmitt (2007) measured reading times for formulaic sequences versus matched non-formulaic phrases in native and non-native speakers of English. The formulaic sequences were read more quickly than the non-formulaic phrases by both groups of participants.

Ellis and Simpson-Vlach (2009) and Ellis et al. (2008) used four experimental procedures to determine how the corpus-linguistic metrics of frequency and mutual information (*MI*, a statistical measure of the coherence of strings) are represented implicitly in native and non-native speakers of English, and how this knowledge affects their accuracy and fluency of processing of the formulas of the Academic Formulas List (AFL, Simpson-Vlach and Ellis 2010, see Section 3.1 for further details). The language-processing tasks in these experiments were selected to sample an ecologically valid range of language-processing skills: spoken and written, production and comprehension, form-focused and meaning-focused. They were: (1) speed of reading and acceptance in a grammaticality judgement task where half of the items were real phrases in English and half were not, (2) rate of reading and rate of spoken articulation, (3) binding and primed pronunciation – the degree to which reading the beginning of the formula primed recognition of its final word, (4) speed of comprehension and acceptance of the formula as being appropriate in a meaningful context. Processing in all experiments was affected by various corpus-derived metrics: length, frequency and mutual information. Frequency was the major determinant for non-native speakers, but for native speakers it was predominantly the *MI* of the formula which determined processability.

Durrant and Schmitt (2009) extracted adjacent English adjective–noun collocations from two learner corpora and two comparable corpora of native student writing and calculated the *t*-score and *MI* score in the *BNC* for each combination extracted. This study also found that non-native

writers rely heavily on high-frequency collocations like *good example* or *long way*, but that they underuse less frequent, strongly associated collocations like *bated breath* or *preconceived notions*. They conclude 'that these findings are consistent with usage-based models of acquisition while accounting for the impression that non-native writing lacks idiomatic phraseology' (2009: 157).

Such findings argue against a clear distinction between linguistic forms that are stored as formulas and ones that are openly constructed. Grammatical and lexical knowledge are not stored or processed in different mental modules, but rather form a continuum from heavily entrenched and conventionalised formulaic units (unique patterns of high token frequency, such as *Hi! How are you?*) to loosely connected but collaborative elements (patterns of high type frequency, such as the generic slot-and-frame pattern *Put [NP] on the table*, which generates a variety of useful tea-time commands: *Put it on the table*, *Put the bread on the table*, *Put the knives and forks on the table*, *Put some plates on the table*, etc.) (Ellis 2008c; Robinson and Ellis 2008; Ellis and Larsen-Freeman 2009; Bybee 2010; Ellis 2012b).

That learners are sensitive to the frequencies of occurrence of constructions and their transitional probabilities suggests that they learn these statistics from usage, tallying them implicitly during each processing episode. Linguistic structure *emerges* from the conspiracy of these experiences (Ellis 1998, 2011). Hopper (1987: 143), in laying the foundations for Emergent Grammar, argued that '[t]he linguist's task is in fact to study the whole range of repetition in discourse, and in doing so to seek out those regularities which promise interest as incipient sub-systems. Structure, then, in this view is not an overarching set of abstract principles, but more a question of a spreading of systematicity from individual words, phrases, and small sets'.

2.2 Three different statistical operationalisations of formulaic language

Section 2.1 argued against a firm distinction between linguistic forms that are stored as formulas and ones that are openly constructed. Instead it proposed that formulaicity is a dimension to be defined in terms of strength of serial dependencies occurring at all levels of granularity and at each transition in a string of forms. At one extreme are formulaic units that are heavily entrenched (high token frequency, unique patterns), at the other are creative constructions consisting of strings of slots each potentially filled by many types. Broadly, the more frequent and the more coherent a string, the faster it is processed. It follows that formulas need to be operationalised in statistical terms that measure frequency and coherence. Statistical operationalisations allow triangulation with corpus samples of the usage which serves as the source of our

knowledge of formulaicity and patterns in language. Corpus-linguistic techniques provide a range of methods for the quantification of recurring sequences (as clusters, n-grams, collocations, phrase-frames, etc.) and for gauging the strength of association between the component words. Three broad options for the basis of determination of formulaic sequences are frequency, association and native norms. Each is considered in turn in the following subsections. (For further studies of phraseological patterning in learner language, see Chapter 10, this volume.)

2.2.1 Frequency

Formulas are recurrent sequences. One definition, then, is that we should identify strings that recur often. This is the approach of Biber and colleagues (Biber et al. 1999; Biber, Conrad and Cortes 2004), who define lexical bundles solely on the basis of frequency. This has the great advantages of being methodologically straightforward and having face validity. We all agree that high-frequency strings like *How are you?*, *Nice day today* and *Good to see you* are formulaic sequences. But we also know some formulas that are not of particularly high frequency, like *blue moon*, *latitude and longitude* and *raining cats and dogs*. And other high-frequency strings, like *and of the* or *but it is*, do not seem very formulaic. Definitions in terms of frequency alone result in long lists of recurrent word sequences that collapse distinctions that intuition would deem relevant. N-grams consisting of high-frequency words occur often. But this does not imply that they have clearly identifiable or distinctive functions or meanings; many of them occur simply by dint of the high frequency of their component words, often grammatical functors. The fact that a formula is above a certain frequency threshold does not necessarily imply either psycholinguistic salience or coherence (Schmitt et al. 2004).

2.2.2 Association

Psycholinguistically salient sequences, on the other hand, like *once in a blue moon*, *on the other hand* or *put it on the table* cohere much more than would be expected by chance. They are 'glued together' and thus measures of association, rather than raw frequency, are more relevant. There are numerous statistical measures of association available, each with their own advantages and disadvantages (Evert 2005; Gries 2008c, 2009, 2012b, 2013d). For example, *MI* is a statistical measure commonly used in information science to assess the degree to which the words in a phrase occur together more frequently than would be expected by chance (Oakes 1998; Manning and Schütze 1999). A higher *MI* score means a stronger association between the words, while a lower score indicates that their co-occurrence is more likely due to chance. *MI* is a scale, not a test of significance, so there is no minimum threshold value; the value of *MI* scores lies in the comparative information they provide. *MI* privileges coherent

strings that are constituted by low-frequency items, like *longitude and latitude*.

2.2.3 Native norms

Definitions purely in terms of frequency or association might well reflect that language production makes use of sequences that are ready made by the speaker or writer, but these need not necessarily be native-like. Non-native academic writing can often be identified by the high frequency of use of phrases that come from strategies of translation from the L1 (mother tongue) (like *make my homework* or *make a diet*), or formulas that occur frequently in spoken language but which are frowned upon as informal in academic writing (like *I would like to talk about* or *I think that…*) (Gilquin and Paquot 2008). An additional, divergent, criterion for formulaicity is that it reflects native-like selection and native-like fluency (Pawley and Syder 1983). Thus we can also operationalise the formulaicity of L2 language by how well it uses the formulaic sequences and grammatico-lexical techniques of the norms of its reference genre. For example, as we will see in Section 3.2, O'Donnell et al. (2013) search for instances of formulaic academic patterns of the AFL (Simpson-Vlach and Ellis 2010) in corpora of native and non-native English academic writing at different levels of proficiency. They show that L2 learners' writing is less rich in the use of these native-norm-derived academic formulas compared to expert native writers.

We are only beginning to explore how these different statistical and corpus-based operationalisations affect acquisition and processing, and this is a research area where much remains to be done. There is strong consensus that research on formulaic language, phraseology and constructions is in dire need of triangulation across research in first and second language acquisition, corpus linguistics, usage-based linguistics and psycholinguistics (Ellis 2008c; Gries 2008c, 2009; Divjak and Gries 2012), and shared operationalisations rest at the foundations of this enterprise.

2.3 L2 learners have difficulty mastering native-like formulaic language

The fields of applied linguistics and SLA showed early interest in multi-word sequences and their potential role in language development. Corder (1973) coined the term *holophrase* to refer to unanalysed multi-word sequences associated with a particular pragmatic function; Brown (1973) called them 'prefabricated routines'. One of the main research questions for SLA researchers at the time was: do prefabricated routines pose a challenge to the traditional view of L1 learning as a process by which children start out with small units (morphemes and words) and then gradually combine them into more complex structures? Do children alternatively and/or additionally start out from large(r) chunks of language which

they then gradually break down into their component parts? Early studies did not yield conclusive results (a good discussion can be found in Krashen and Scarcella 1978). For example, Hakuta (1976), based on data from a 5-year-old Japanese learner of English, argued in favour of a more fine-grained distinction between prefabricated routines and prefabricated patterns, that is, low-scope patterns that have at least one variable slot. Wong Fillmore's (1976) dissertation project was one of the first to track more than one child over a longer period of time; her analysis suggested that ESL (English as a Second Language) children do in fact start out with prefabricated patterns which they gradually break down into their component parts in search of the rules governing their L2, which, in turn, ultimately enables them to use language creatively.

There were only a few early studies on adult L2 learners (Wray 2002: 172–98 provides a detailed overview). The general consensus, however, was that while adult L2 learners may occasionally employ prefabricated language, there was less evidence than in children's data that knowledge of prefabricated language would foster grammatical development in adult L2 acquisition (L2A). Hanania and Gradman (1977), for instance, studied Fatmah, a native speaker of Arabic. Fatmah was 19 years old at the time of the study, and she had received little formal education in her native language. When speaking English, Fatmah used several routines that were tied to specific pragmatic situations; however, the researchers found her largely unable to analyse these routines into their component parts. Similarly, Schumann (1978), who investigated data from several adult L2 learners with different native language backgrounds, found little evidence in favour of prefabricated language use. A slightly different picture emerged in Schmidt's (1983) well-known research on Wes, a native speaker of Japanese who immigrated to Hawaii in his early thirties. Wes seemed to make extensive use of prefabricated routines. However, while this significantly boosted Wes's fluency, his grammatical competence remained low. Ellis (1984), looking at the use of prefabricated language in an instructional setting, suggested that there is considerable individual variation in learners' ability to make the leap from prefabricated routines to the underlying grammatical rules they exemplify. Krashen and Scarcella (1978) were outright pessimistic regarding adult learners' ability to even retain prefabricated routines, and cautioned against focusing adult learners' attention on prefabricated language because '[t]he outside world for adults is nowhere near as predictable as the linguistic environment around Wong Fillmore's children was' (Krashen and Scarcella 1978: 298).

In their classic analysis of formulaic language usage in SLA, 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency', Pawley and Syder (1983) put the clear case that L2 speakers, despite considerable knowledge of L2 grammar, still make productions that are unidiomatic. Likewise, in her analysis of the incidence of formulaic language in French

students' advanced EFL (English as a Foreign Language) writing, Granger (1998c) showed that learners made less use of formulaic expressions and collocations than native writers.

The studies reviewed here suggest a potential difference in formulaic use between ESL learners who are exposed to large amounts of naturalistic spoken language and EFL learners who are not. Learning the usages that are normal or unmarked from those that are unnatural or marked requires a huge amount of immersion in the speech community. Language learning is essentially a sampling problem – the learner has to estimate the native norms from a sample of usage experience (Ellis 2008b). Many of the forms required for idiomatic use are of relatively low frequency, and the learner thus needs a large input sample just to encounter them:

> Becoming idiomatic and fluent requires a sufficient sample of needs-relevant authentic input for the necessary implicit tunings to take place. The 'two puzzles for linguistic theory', nativelike selection and nativelike fluency (Pawley and Syder, 1983), are less perplexing when considered in these terms of frequency and probability. There is a lot of tallying to be done here. The necessary sample is certainly to be counted in terms of thousands of hours on task. (Ellis 2008b: 152)

2.4 L2 longitudinal research: from formula to low-scope pattern to creative construction?

That L2 learners have difficulty in acquiring the full range of native-like formulaic expressions does not mean that some high-frequency formulas do not play a part in language acquisition. There are recent longitudinal studies in support of this developmental sequence. Particular formulas, high in frequency, functionality and prototypicality might serve as pacemakers.

Myles and colleagues (Myles et al. 1998; Myles et al. 1999; Myles 2004) analysed longitudinal corpora of oral language in secondary school pupils learning French as a foreign language in England. The study investigated the development of chunks within individual learners over time, showing a clear correlation between chunk use and linguistic development:

> In the beginners' corpus, at one extreme, we had learners who failed to memorise chunks after the first round of elicitation; these were also the learners whose interlanguage remained primarily verbless, and who needed extensive help in carrying out the tasks. At the other extreme, we had learners whose linguistic development was most advanced by the end of the study. These were also the learners who, far from discarding chunks, were seen to be actively working on them throughout the data-collection period. These chunks seem to provide these learners with a databank of complex structures beyond their current grammar,

> which they keep working on until they can make their current generative grammar compatible with them.
>
> (Myles 2004: 153)

This study is such a landmark that we have chosen it for further detailed examination in Section 3.3.

Eskildsen and Cadierno (2007) investigated the development of *do*-negation by a Mexican learner of English. *Do*-negation learning was found to be initially reliant on one specific instantiation of the pattern *I don't know*, which thereafter gradually expanded to be used with other verbs and pronouns as the underlying knowledge seemed to become increasingly abstract, as reflected in token and type frequencies. The emerging system was initially based on formulaic sequences, and development was based on the gradual abstraction of regularities that link expressions as constructions (see also Eskildsen 2012).

Mellow (2008) describes a longitudinal case study of a 12-year-old Spanish learner of English, Ana, who wrote stories describing fifteen different wordless picture books during a 201-day period. The findings indicate that Ana began by producing only a few types of complex constructions that were lexically selected by a small set of verbs, which gradually then seeded an increasingly large range of constructions.

Sugaya and Shirai (2009) describe acquisition of Japanese tense–aspect morphology in L1 Russian learner Alla. In her ten-month longitudinal data, some verbs (e.g. *siru* 'come to know', *tuku* 'be attached') were produced exclusively with imperfective aspect marker *-te i-(ru)*, while other verbs (e.g. *iku* 'go', *tigau* 'differ') were rarely used with *-te i-(ru)*. Even though these verbs can be used in any of the four basic forms, Alla demonstrated a very strong verb-specific preference. Sugaya and Shirai follow this up with a larger cross-sectional study of sixty-one intermediate and advanced learners who were divided into thirty-four lower- and twenty-seven higher-proficiency groups using grammaticality judgement tasks. The lower-proficiency learners used the individual verbs in verb-specific ways and this tendency was stronger for the verbs denoting resultative state meaning with *-te i-(ru)* (e.g. achievement verbs) than the verbs denoting progressive meaning with *-te i-(ru)* (e.g. activity, accomplishment verbs). Sugaya and Shirai conclude that learners begin with item-based learning and 'low-scope patterns' and that these formulas allow them to gradually gain control over tense–aspect. Nevertheless, they also consider how memory-based and rule-based processes might co-exist for particular linguistic forms, and how linguistic knowledge should be considered a 'formulaic–creative continuum'.

Having said that, there are studies of L2 that have set out to look for the developmental sequence from formula to low-scope pattern to creative construction in a learner corpus and found less compelling evidence. These are reviewed below.

Bardovi-Harlig (2002) studied the emergence of future expressions involving *will* and *going to* in a longitudinal corpus study of sixteen adult ESL learners (mean length of observation: 11.5 months; 1,576 written texts, mainly journal entries, and 175 oral texts, either guided conversational interviews or elicited narratives based on silent films). The data showed that future *will* emerges first and greatly outnumbers the use of tokens of *going to*. Bardovi-Harlig (2002: 192) describes how the rapid spread of *will* to a variety of verbs suggests that 'for most learners, there is either little initial formulaic use of *will* or that it is so brief that it cannot be detected in this corpus'. There was some evidence of formulaicity in early use of *going to*: '[f]or 5 of the 16 learners, the use of *I am going to write* stands out. Their productions over the months of observation show that the formula breaks down into smaller parts, from the full *I am going to write about* to the core *going to* where not only the verb but also person and number vary. This seems to be an example of learner production moving along the formulaic–creative continuum' (2002: 197). But other learners showed greater variety of use of *going to*, with different verbs and different person-number forms, from its earliest appearance in the diary. Bardovi-Harlig (2002: 198) concludes that 'although the use of formulaic language seems to play a limited role in the expression of future, its influence is noteworthy'.

Eskildsen (2009) analysed longitudinal oral second language classroom interaction for the use of *can* by one student, Carlo. *Can* first appeared in the data in the formula *I can write*. But Eskildsen noted how formulas are interactionally and locally contextualised, which means that they may possibly be transitory in nature, their deployment over time being occasioned by specific recurring usage events.

Hall (2010) reports a small-scale study of the oral production of three adult beginner learners of ESL over a nine-week period in a community language programme meeting three days per week for two hours each day. A wide variety of tasks was used to elicit the data, which included picture description and semi-structured interviews. Hall reports that formulas were minimally present in the learner output and that constructions and formulas of similar structure co-existed, but that a developmental relationship between formulas and constructions was not clearly evident. He concludes that the amount of elicited data was too limited to substantiate the learning path under investigation, and that more controlled task dimensions were also needed.

3 Representative studies

We have chosen four research studies to illustrate a range of different approaches to these issues. The first identifies formulas from corpora of genre-specific language and then assesses L1 and L2 knowledge of these

formulas in psycholinguistic experiments. The second uses cross-sectional learner corpora to investigate the development of formulaic language in first and second language writing, investigating effects of statistical operationalisation in terms of frequency, association and native norm. The third is a mixed-methods longitudinal corpus-plus-experimentation study of the role of formulas in language learning in secondary school. The fourth tracks constructions over time in a longitudinal corpus of naturalistic second language acquisition in adults, investigating type–token frequency distributions in verb-argument constructions over time, the ways in which native speaker usage guides learner language, how constructions develop following psychological principles of category learning, and complementing observational description with computational simulations.

3.1 Simpson-Vlach, R. C. and Ellis, N. C. 2010. 'An academic formulas list: New methods in phraseological research', *Applied Linguistics* 31(4): 487–512.

Our first representative study is not a learner corpus study per se, but one which uses corpus techniques to identify the formulas in academic language so that learner knowledge of these could then be evaluated firstly by using psycholinguistic approaches (Ellis et al. 2008; Ellis and Simpson-Vlach 2009) and secondly by searching for these expressions in learner corpora (O'Donnell et al. 2013).

Simpson-Vlach and Ellis (2010) used corpus-linguistic techniques to identify the phraseology specific to academic discourse. The resultant Academic Formulas List includes formulaic sequences identified as (1) frequent recurrent patterns in corpora of written and spoken language, which (2) occur significantly more often in academic than in non-academic discourse, and (3) inhabit a wide range of academic genres. Three-, four- and five-word formulas occurring at least ten times per million words were extracted from corpora of 2.1 million words of academic spoken language [*Michigan Corpus of Academic Spoken English, MICASE*,[3] and selected academic spoken *BNC* files], 2.1 million words of academic written language [Hyland's (2004a) research article corpus, plus selected academic writing *BNC* files], 2.9 million words of non-academic speech [the *Switchboard*[4] corpus] and 1.9 million words of non-academic writing [the *FLOB*[5] and *Frown*[6] corpora gathered in 1991 to reflect British and American English over fifteen genres]. The program *Collocate* (Barlow 2004) allowed the authors to measure the frequency of each n-gram along with the *MI* score for each phrase.

[3] http://quod.lib.umich.edu/m/micase/ (last accessed on 13 April 2015).

[4] https://catalog.ldc.upenn.edu/LDC97S62 (last accessed on 13 April 2015).

[5] http://clu.uni.no/icame/manuals/FLOB/INDEX.HTM (last accessed on 13 April 2015).

[6] http://clu.uni.no/icame/manuals/FROWN/INDEX.HTM (last accessed on 13 April 2015).

The total number of formulas appearing in any one of the four varieties at the threshold level of ten per million words was approximately 14,000. In order to determine which formulas were more frequent in the academic corpora than in their non-academic counterparts, the authors used the log-likelihood (LL) statistic (Oakes 1998) to determine the formulas which were statistically more frequent, at a significance level of $p < 0.01$. They separately compared academic speech vs non-academic speech, resulting in over 2,000 items, and academic writing vs non-academic writing, resulting in just under 2,000 items. There was also a smaller core list of formulas that are common in academic spoken *and* academic written language.

Simpson-Vlach and Ellis (2010: 496) then took a stratified sample of these formulas and asked experienced English for Academic Purposes (EAP) instructors and language testers to rate them on three judgement scales for: (a) whether or not they thought the phrase constituted 'a formulaic expression, or fixed phrase, or chunk'; (b) whether or not they thought the phrase had 'a cohesive meaning or function, as a phrase'; or (c) whether or not they thought the phrase was 'worth teaching, as a bona fide phrase or expression'. This allowed the authors to further prioritise these items in the AFL using an empirically derived measure of utility that is educationally and psychologically valid and operationalisable with corpus-linguistic metrics.

Simpson-Vlach and Ellis (2010) present the AFL formulas according to their predominant pragmatic functions, including, for example, hedges (e.g. *there may be, to some extent*, *you might want to*), evaluative expressions (*the importance of*, *is consistent with*, *it is obvious that*, *it doesn't matter*), and textual reference (*in the next section*, *shown in table*). These categories illustrate the nature of academic language and they guide the use of the AFL in EAP instruction.

Ellis et al. (2008) and Ellis and Simpson-Vlach (2009) researched these formulas in psycholinguistic experiments showing that different aspects of formulaicity affect the accuracy and fluency of language processing in native speakers and in advanced L2 learners of English (details are given in Section 2.1). For native speakers, it was predominantly the *MI* of the formula which determined processability, for L2 learners of the language, it was predominantly the frequency of the formula. These findings inform usage-based theories of language learning and processing (Ellis 2012b).

3.2 O'Donnell, M. B., Römer, U. and Ellis, N. C. 2013. 'The development of formulaic language in first and second language writing: Investigating effects of frequency, association, and native norm', *International Journal of Corpus Linguistics* 18: 83–108.

Replicable research must be grounded upon operational definitions in statistical terms. However, there is variability in the research literature over the definition and measurement of formulaic sequences, as well as in methods of corpus comparison. O'Donnell et al. (2013) therefore adopted

an experimental design and applied four different corpus-analytic measures, variously based upon n-gram frequency (frequency-grams), association (MI-grams), phrase-frames (p-frames, see Fletcher 2002–2007) and native norms (the AFL items described above in Section 3.1 – AFL-grams), to samples of English writing produced by native speakers and by second/foreign learners of different first language backgrounds (Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish and Swedish) in order to examine and compare knowledge of formulas in first and second language acquisition as a function of proficiency and language background.

Corpora of writing were sampled from different L1 backgrounds and at a range of proficiency levels, including: European University English learner writing (*International Corpus of Learner English, ICLE*; Granger et al. 2002), undergraduate native English student writing (*Louvain Corpus of Native English Essays, LOCNESS*),[7] A-graded graduate writing by non-native English speakers (*Michigan Corpus of Upper-level Student Papers, MICUSP-NNS*; Römer and O'Donnell 2011), A-graded final year undergraduate and graduate writing by native English speakers (*MICUSP-NS*; Römer and O'Donnell 2011), and the Hyland collection of published research articles written by native or near-native English speakers (Hyland 1998, 2004a). O'Donnell et al. (2013) took eight independent random samples from each of these corpora and quantified and compared the frequencies and learner uptake of continuous sequences of various lengths (n-grams, e.g. *at the end of*) and associated 'frames' (e.g. *at the * of*). Various statistical analyses were applied to investigate the effects of (a) proficiency development in the usage of these units and (b) L1 backgrounds.

The different operationalisations produced different patterns of effect of expertise and L1/L2 status.

For frequency-defined formulas, there were effects of expertise (Expert ≈ Graduate > Undergraduate), with, if anything, L2 learners producing more formulas than their native peers. O'Donnell et al. suggest that these are likely effects of text sampling on the recurrence of formulaic patterns, with the prompt questions driving the more common formulaic sequences in *ICLE* (e.g. *the opium of the masses*, *the birth of a nation*, *the generation gap*, *ICLE French*) and *LOCNESS* (e.g. *the Joy Luck Club*, *in Le Myth de Sysiphe*, *the root of all evil*). *MICUSP* (especially *MICUSP-NS*) generated common formulaic sequences from reference sections (e.g. *American Journal of Public Health*, *Hispanic Journal of Behavioral Sciences*, *levels of psychological well-being*). The Hyland corpus, with its greater diversity of topics across disciplines, showed less of these sampling foci.

For AFL-defined formulas, there were clear effects of high levels of expertise (Expert > A-grade Graduate ≈ Undergraduate), but no effect of L1/L2 status. The expert (Hyland corpus) authors were senior scholars, who had had multiple-year university training and experience in getting

[7] www.uclouvain.be/en-cecl-locness.html (last accessed on 13 April 2015).

published in peer-reviewed journals. They were clearly differentiated from both the novice academic writers who contributed to *ICLE* and *LOCNESS*, and those who produced A-grade *MICUSP* papers on their way to developing expert writing skills and becoming accepted members of academic communities of practice. The fact that there were no effects of L1/L2 status suggests that these means of expression are as novel and specialised for natives as for non-natives.

These analyses thus show clear effects of operationalisation of 'formulaic language' and of the choices underlying the design of different corpora. We will consider the implications further in Section 4.

3.3 Myles, F., Mitchell, R. and Hooper, J. 1999. 'Interrogative chunks in French L2: A basis for creative construction?', *Studies in Second Language Acquisition* 21(1): 49–80.

In an extensive study of secondary school pupils learning French as a foreign language in England, Myles (Myles et al. 1998; Myles et al. 1999; Myles 2004) analysed longitudinal corpora of oral language in sixteen Beginners [Years 7, 8 and 9 (11–14 years old), tracked over the first 2¼ years, using thirteen oral tasks (2–3 per term over six terms)] and sixty Intermediates [20 classroom learners in each of Years 9, 10 and 11 studied cross-sectionally using four oral tasks (three repeated from the Beginners project)]. These data showed that multimorphemic sequences which go well beyond learners' grammatical competence are very common in early L2 production. Notwithstanding that these sequences contain such forms as finite verbs, *wh*-questions and clitics, Myles denies this as evidence for the sequences being openly created by syntactic means from the start of L2 acquisition because the relevant functional projections were not present outside chunks initially. Analyses of inflected verb forms suggested that early productions containing them were formulaic chunks. These structures, sometimes highly complex syntactically (e.g. in the case of interrogatives), cohabited for extended periods of time with very simple sentences, usually verbless or, when a verb was present, normally untensed. Likewise, clitics first appeared in chunks containing tensed verbs, suggesting that it is through these chunks that learners acquire them. Myles characterises these early grammars as consisting of lexical projections and formulaic sequences, showing no evidence of open syntactic creation. 'Chunks do not become discarded; they remain grammatically advanced until the grammar catches up, and it is this process of resolving the tension between these grammatically advanced chunks and the current grammar which drives the learning process forward' (Myles 2004: 152).

The results of this extensive corpus study were reported in three or four papers, each concentrating on different linguistic constructions. Myles's conclusion for the relationship between formulaic chunks and creative construction was not that the direction of development was one

of integration (from words to formulas) or one of differentiation (from formulaic phrases to their components), but rather that 'creative construction and chunk breakdown clearly go hand in hand' (Myles et al. 1999: 76):

> We see, on the one hand, chunks becoming simpler and more like other constructions present in the grammar at a given time and, on the other hand, creative constructions becoming more complex as elements from the chunks feed into the process. It is as if, at any one time, learners are attempting to resolve the tension between complex but communicatively rich chunks on the one hand and simple but communicatively inadequate structures on the other hand. This is a dynamic tension that drives forward the overall development of the L2 system. (Myles et al. 1999: 77)

3.4 Ellis, N. C. and Ferreira-Junior, F. 2009a. 'Constructions and their acquisition: Islands and the distinctiveness of their occupancy', *Annual Review of Cognitive Linguistics* 7: 187–220.

Ellis, N. C. and Ferreira-Junior, F. 2009b. 'Construction learning as a function of frequency, frequency distribution, and function', *The Modern Language Journal* 93: 370–86.

Ellis and Ferreira-Junior (2009a, 2009b) were interested in the processes of integration and differentiation of formulaic and semi-formulaic phrases in the acquisition of more schematic constructions in naturalistic second language acquisition. They therefore investigated effects of type–token distributions in the slots comprising the linguistic form of three English verb-argument constructions (VACs), namely verb locative (VL), e.g. *Tom walked to the store*, verb object locative (VOL), e.g. *he put the book on the shelf* and ditransitive (VOO), e.g. *he sent his son some money*, in the speech of second language learners in the *European Science Foundation* (*ESF*) corpus (Feldweg 1991; Perdue 1993; Dietrich et al. 1995). The *ESF* project collected the spontaneous and elicited second language of adult immigrants recorded longitudinally in interviews every four to six weeks for approximately thirty months. Ellis and Ferreira-Junior focused upon seven ESL learners living in Britain whose native languages were Italian (n=4) or Punjabi (n=3). The *ESF* corpus includes transcribed data from 234 sessions for these ESL learners and their native-speaker conversation partners during a range of activities.

Goldberg (2006) had previously argued for child language acquisition that Zipfian[8] (Zipf 1935) type–token frequency distribution of verbs in natural language might optimise construction learning by providing

[8] In natural language, Zipf's (1935) law describes how the highest frequency words account for the most linguistic tokens. The frequency of words decreases as a power function of their rank in the frequency table, with the most frequent word occurring approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc.

one very high-frequency exemplar that is also prototypical in meaning. Ellis and Ferreira-Junior (2009b) confirmed that, in the naturalistic L2A of English, VAC verb type–token distribution in the input is Zipfian and learners first acquire the most frequent, prototypical and generic exemplar (e.g. *go* in VL, *put* in VOL, *give* in VOO). Ellis and Ferreira-Junior (2009a) further illustrate how acquisition is affected by the frequency and frequency distribution of exemplars within each island of the construction (e.g. [Subj V Obj $Obl_{PATH/LOC}$]), by their prototypicality, and, using a variety of psychological and corpus-linguistic association metrics, by their contingency of form–function mapping and by the degree to which the different elements in the VAC sequence (such as Subj V Obj Obl) are mutually informative and form predictable chunks. The highest-frequency elements seeding the learners' VL pattern were *go to the shop*, the VOL pattern *put it on the table* and the VOO pattern *they give me money*. We will describe in more detail in Section 4 the cycles of integration and differentiation whereby overlapping chunks of formulaic phrases resonate with creative constructions. Ellis and Larsen-Freeman (2009) used computational (emergent connectionist) models to test theories of how these various factors play out in the emergence of constructions as generalised linguistic schemas from the *ESF* learners' analysis of patterns in their usage history.

4 Critical assessment and future directions

4.1 Corpus design and formulaicity

The research reviewed above allows us to identify aspects of corpus design which affect the incidence of formulaicity and which inform the design and analysis of future studies.

1. There are several well-justified but divergent operational definitions of formulaicity. Choices of operationalisation entail that different researchers are potentially researching and theorising different phenomena.
2. Formulaicity may vary as a function of first vs second language acquisition. L1 acquisition (L1A) may indeed be more formulaic than L2A. When child L1 learners are learning about language from formulaic frames (Mintz 2003; Tomasello 2003; Ambridge and Lieven 2011) and the analysis of sequences of words (Kiss 1973; Elman 1990; Redington and Chater 1998), they are learning from scratch about more abstract categories such as verb, pronoun, preposition, noun or transitive frame. It is debatable whether the units of early L1A are words at all (Peters 1983). Adult L2 learners already know about the existence of these units, categories and linguistic structures. They expect that there will be words and constructions in the L2 which correspond to such word classes and frames. Once they have identified them, or

even, once they have searched them out and actively learned such key vocabulary, they are more likely therefore to attempt creative construction, swopping these elements into corresponding slots in frames. Transfer from the L1 is also likely to affect the process (Granger 1998c; Chapter 15, this volume). The more learners attempt word-by-word translation from their L1, the more they deviate from L2 idiomaticity. There is unconscious transfer too (Jiang and Nekrasova 2007).

3. The amount and type of language exposure is influential (e.g. ESL vs EFL) (Groom 2009; Reppen 2009). Children are naturalistic language learners from thousands of hours of interaction and input. While some adults learn naturalistically, others take grammar-rich courses and foreign language environments provide only restricted access to authentic language. Thus second language can be more formulaic than foreign language.
4. For studies that seek to trace the development of formulaic language, data has to be dense enough to identify repeated uses at the time of emergence (Tomasello and Stahl 2004). The use of formulas and constructions is determined by context, function, genre and register. If elicitation tasks vary, the chance of sampling the same formula and its potential variants diminishes accordingly. Myles (2004) demonstrates that an understanding of L2A can only come from analysis of extensive representative corpora of language sampled in the same learners over time. This, with transcription, mark-up, checking and distribution, entails huge effort. Myles also illustrates how supplementing the language data with targeted psycholinguistic experimental tasks, focused upon times of critical change, can enhance the value of the corpus description.

   The field of child language acquisition became a scientific enterprise upon the recognition of the need for proper longitudinal corpora describing individual language development (Brown 1973). More recently, this has become recognised as a need for dense longitudinal corpora of naturalistic language development that capture perhaps 10 per cent of the child's speech and the input they are exposed to, collected from 2–4 years old when the child is undergoing maximal language development (Maslen et al. 2004; Behrens, 2008), or even a complete corpus of a learner's situated language development (Roy 2009). The making available of the evidence of learner language through *CHILDES* and *TalkBank* (MacWhinney 2000) has transformed the study of child language acquisition. Although beginnings have been made for L2, for example the *ESF* longitudinal corpora (Klein and Perdue 1992), we must together strive for a similar richness of evidential sources for SLA research too (Ortega and Iberri-Shea 2005).
5. As in all other areas of language processing, recognition of formulas is easier than production. Ellis and Ferreira-Junior (2009a, 2009b) showed that naturalistic adult L2 learners used the same verbs in

frequent verb-argument constructions as are found in their input experience, with the relative ordering of the types in the input predicting uptake with correlations in excess of $r = 0.90$. Nevertheless, while they would accurately produce short simple formulaic sequences such as *come in* or *I went to the shop*, structurally more complex constructions were often produced incorrectly. Thus psycholinguistic studies of formula recognition may identify wider knowledge than is evidenced in formula production in learner corpora.

6. Modality, genre and task are also important. Using the range of methods of O'Donnell et al. (2013) described in Section 3.2, Ellis et al. (2009) showed that oral language was much denser in formulaic language than was written news reporting or light fiction. Likewise, the greater the working-memory demands of the processing task, the greater the need to rely on formulas: Kuiper (1996) analysed 'smooth talkers' – sports commentators and auctioneers who are in communicative contexts which place pressure to observe what is transpiring around them, analyse these happenings in short-term memory and formulate speech reports describing what is observed in real time without getting left behind. Smooth talkers use many formulas in their speech – recurrent sequences of verbal behaviour, whether conventional or idiosyncratic, which are sequentially and hierarchically organised. The faster the action, the more difficult it is for the commentator to provide an instantaneous commentary. By contrasting fast-action commentators (horse races, antique and livestock auctioneers) with slow-action commentators (cricket, real estate auctioneers), Kuiper showed that the fast-action commentators made much more use of formulas than did the slow-action ones. We expect similar resort to formulaic language whenever L1 or L2 language users have to speak under conditions of high cognitive demand.
7. Corpus design features including the number of participants, the nature of their tasks and prompts, the amount of language they produce, etc., are potent determinants of outcome. There remains much basic research to be done to assess how formulaicity is affected by potential independent variables of concern for control purposes (text length, type–token ratio, mean length of utterance, entropy, vocabulary frequency profiles, number of speakers, range of prompts and topics, etc.) and by variables of greater theoretical weight, including potential text variables such as spoken/written genre, potential subject variables such as native vs second language status, proficiency, education, and potential situational variables such as degree of preparation, rehearsal and working-memory demand.
8. With so many variables in play in the emergence of linguistic constructions and system (Ellis 2011), an essential part of testing theories of development includes their investigation using computational models as applied to learner corpus data (see further, Ellis 2012b).

4.2 The roles of formulaic language in SLA

The evidence reviewed above demonstrates that (1) language learners have substantial statistical knowledge of the units of language and their phraseological patterning; (2) when one compares second/foreign language to first language, the former displays a smaller range of formulaic expressions; (3) formulaic language can serve in the language acquisition process. Let us bring these three facts together.

Some formulaic sequences are readily learnable by dint of being highly frequent and prototypical in their functionality – *How are you?*, *It's lunchtime*, *I don't know*, *Good example*, *I am going to write about…* and the like. These are good candidates for construction seeds.

Other formulaic sequences are not readily learnable – these are of low frequency, often indeed rare, and many are non-transparent and idiomatic in their interpretation (e.g. *once in a blue moon*, *bated breath*). As idioms they must be learned as such. However, learners require considerable language experience before they encounter these once, never mind sufficient times to commit them to memory (Ellis 2008b; Ellis et al. 2008). This is why learners typically do not achieve native-like idiomaticity (Pawley and Syder 1983; Granger 1998b; Durrant and Schmitt 2009). These low-frequency, low-transparency formulas are targets for learning rather than seeds of learning.

In the huge middle ground between high and low token-frequency formulaic expressions, there is interaction. Let us consider this 'formulaic–creative continuum' (Sugaya and Shirai 2009: 440), the 'repeated cycles of integration and differentiation' (Studdert-Kennedy 1991: 25) or the 'dynamic tension that drives forward the overall development' (Myles et al. 1999: 77) in further detail, with the aid of a corpus, of course.

Begin with the formula (i) *put it in*, and put it in its context of usage in a large corpus of English, such as *COCA*: *put it in* occurs 3,620 times.[9] Consider it as a formulaic exemplification of the schematic verb-object-locative (VOL) verb-argument construction (VAC) which can describe a routine generic caused-motion function of moving something to a new place or in a new direction. Compare it to other VOL VACs. Search for *put it [i*]*.[10] This is very common (8,065 token occurrences), from *put it in* (3,620), *put it on* (1,926), *put it onto* (745) (all highly functional, stereotypical, formulaic phrases in their own right), with then the distribution dropping rapidly to a heavy right tail of items that appear just once, such as *put it away*. These frequencies broadly follow a Zipfian distribution (Zipf 1935; Solé et al. 2005; Ninio 2011; see footnote 8), as in language overall, but not following the particular ordering found in language as a whole – each slot attracts particular types of occupants (Ellis and O'Donnell 2012). A learner

[9] Numbers may differ because the corpus is always growing.

[10] This is the wildcard for any preposition.

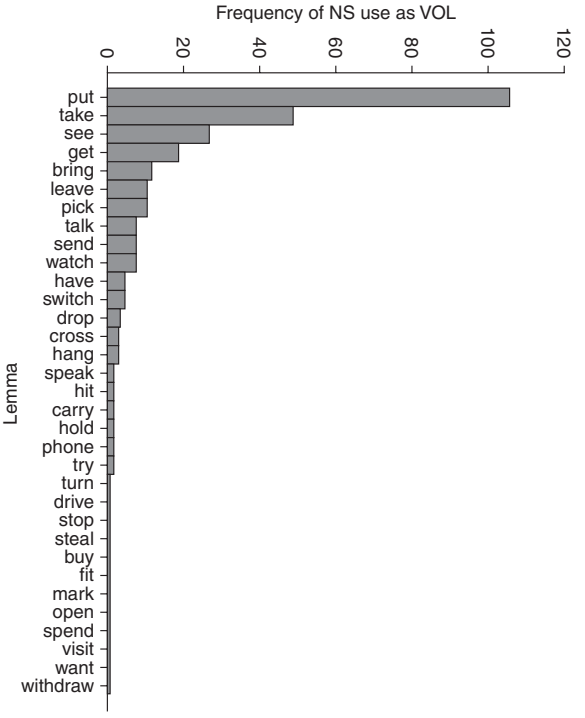

Figure 16.1 The Zipfian type–token frequency distribution of verb lemmas in the VOL VAC in the native English participants of the *ESF* project (based on Ellis and Ferreira-Junior 2009a)

would get a very good idea of locatives by abstracting over these types and tokens of prepositions.

Next consider the types of verbs that work in these constructions. Searching *[v*]*[11] *it [i*]* produces *put it in* (3,608), *give it to* (2,521), *do it in* (2,059), *put it on* (1,917) (again all formulaic)… There are many more types here but the frequencies still follow a Zipfian distribution. Figure 16.1 shows the results of a parallel analysis of the verb types in VOL constructions from the native English speakers in the *ESF* corpus from Ellis and Ferreira-Junior (2009a). There is some noise, but abstracting over the verb types, of which *put* takes the lion's share in useful, stereotypically functional formulaic phrases such as *put it in*, *put it on*, *put it onto*, the learner would get a pretty good idea of the semantics of caused-motion verbs.

Back to *COCA*, a more specific search with *put it in the* * generates *put it in the oven* (53), *put it in the refrigerator* (28), *put it in the back* (27), *put it in the freezer* (26),… *put it in the hold* (2). The sorts of everyday places where people put things in are pretty clear in their semantics too, when averaged thus. And who puts? Searching *[p*]/[n*]*[12] *put it* generates *you put it*

[11] This is the wildcard for any verb.

[12] This is the wildcard for any pronoun or noun.

(1,067), *he put it* (975), *I put it* (891), … *who put it* (72), *official put it* (62), etc. The learner would get a clear idea of the sorts of entities who do the putting. There are exceptions, but there is semantic coherence over the general exemplar cloud.

In each of these analyses there is a broadly Zipfian type–token frequency distribution within the slot; the most frequent, pathbreaking slot-filler for each VAC is much more frequent than the other members; the most frequent slot-filler is semantically prototypical and generic of the VAC island as a whole.

This analysis in *COCA* was seeded with a frequent formulaic prototype VOL, *put it in*, with its characteristic form and its generic interpretation. Scrutiny of its component slots and the types they attract in usage generated other VOLs with high-frequency prototypical occupants. Abstracting over the typical types in the various slots results in a generalised schema for the VOL, with the different slots becoming progressively defined as attractors. Each slot in each construction thus makes a significant contribution to its identification and interpretation (Tomasello 2003; Goldberg 2006; Ellis and Ferreira-Junior 2009a, 2009b; Ellis and Larsen-Freeman 2009; Bybee 2010; Ambridge and Lieven 2011; Ellis and O'Donnell 2012).

Is the notion of language acquisition being seeded by formulaic phrases and yet learner language being formula-light illogical? Is this 'having your cake and eating it too'? Pawley and Syder (1983) thought not. While much of their classic article concentrated on the difficulty L2 learners had in achieving native-like formulaic selection and native-like fluency, nevertheless they stated '[i]ndeed, we believe that memorized sentences are the normal building blocks of fluent spoken discourse, and at the same time, that they provide models for the creation of many (partly) new sequences which are memorable and in their turn enter into the stock of familiar uses' (1983: 208). Granger's (1998c) analysis of collocations and formulas in advanced EFL writing showed likewise that 'learners use fewer prefabs than their native-speaker counterparts' while at the same time they use some lexical teddy bears as 'general-purpose amplifiers' in booster and maximiser phrases – 'the analysis showed a highly significant overuse of *very* as the all-round amplifier par excellence … one could postulate that the learners' underuse of *ly* amplifiers is compensated for by their overuse of *very*' (1998c: 151). At this stage of learning, *very [adj]* is the 'all-round amplifier par excellence', the memorised and prototypical model of amplifier phrases yet to come.

The present characterisation of the developmental sequence 'from formula to low-scope pattern to creative construction' is less true to the traditional idea of a formula as categorically defined, and more so to that of formulaicity as a variable reflecting sequential dependencies in usage and degree of entrenchment in the learner's mind. To properly investigate these questions, we need more longitudinal studies based on dense data (see also Chapter 17, this volume), more studies that

compare formulaic language in L1 vs L2, more studies that compare formulaic language development in second vs foreign language acquisition, and more studies that compare formulaic language in recognition vs comprehension. Only then will we be able to put rich, quantitative flesh on the core, skeletal claim that 'grammar is what results when formulas are re-arranged, or dismantled and re-assembled, in different ways' (Hopper 1987: 145).

Key readings

Robinson, P. and Ellis, N. C. (eds.) 2008. *Handbook of Cognitive Linguistics and Second Language Acquisition*. London: Routledge.

This edited collection brings together leading researchers in usage-based first and second language acquisition. Usage-based approaches hold that we learn language from our experience of language, and that formulaic language plays a key role. This is the first volume to extend cognitive-linguistic analyses across L1A and L2A.

Polio, C. (ed.) 2012. *Topics in Formulaic Language.* Special issue of *Annual Review of Applied Linguistics* 32.

This is a recent, comprehensive, and broad-ranging collection of twelve articles reviewing cognitive perspectives in L1A, L2A, language processing, language disorders, formulaic language pedagogy, and social perspectives on formulaic language.

Paquot, M. and Granger, S. 2012. 'Formulaic language in learner corpora', *Annual Review of Applied Linguistics* 32: 130–49.

This is a recent state-of-the-art review of learner corpus studies.

Rebuschat, P. and Williams, J. N. (eds.) 2012. *Statistical Learning and Language Acquisition*. Berlin: Mouton de Gruyter.

Linguistic constructions are acquired from experience of input following associative learning principles. This collection on statistical language learning considers theories of how type–token frequency patterns in the input, patterns that can only be ascertained from corpus analysis, drive the statistical learning that results in categorisation.

Hoffmann, Th. and Trousdale, G. (eds.) 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press.

This is a recent collection on construction grammar and usage-based acquisition. A central theme is the interplay between formulaic language and more open constructions and their synergy in language acquisition, knowledge, processing, and change.