LANGUAGE LEARNING
A Journal of Research in Language Studies

Cognition, Corpora, and Computing: Triangulating Research in Usage-based Language Learning

Journal:	*Language Learning*
Manuscript ID	16-CRA-1816.R2
Wiley - Manuscript type:	Conceptual Review Article
Keywords:	Cognition, Corpora, Language complexity, Language acquisition, Usage-based approaches, Psycholinguistics

Abstract:	Usage-based approaches explore how we learn language from our experience of language. Related research thus involves the analysis of the usage from which learners learn, and of learner usage as it develops. This program involves considerable data recording, transcription, and analysis, using a variety of corpus and computational techniques, many of them specially devised for learner language. This paper surveys relevant developments across the psychology of learning, first and second language acquisition, psycholinguistics, corpus linguistics, and computational linguistics, and identifies challenges and future priorities relating to the following issues: (1) analyzing the distributional characteristics of linguistic constructions and their meanings in large collections of language that are representative of the language that learners experience, (2) the longitudinal analysis of learner language, (3) Natural Language Processing (NLP) analyses of the dimensions of language complexity.

Cognition, Corpora, and Computing:
Triangulating Research in Usage-based Language Learning

N. C. Ellis
ncellis@umich.edu
University of Michigan

Abstract

Usage-based approaches explore how we learn language from our experience of language. Related research thus involves the analysis of the usage from which learners learn, and of learner usage as it develops. This program involves considerable data recording, transcription, and analysis, using a variety of corpus and computational techniques, many of them specially devised for learner language. This paper surveys relevant developments across the psychology of learning, first and second language acquisition, psycholinguistics, corpus linguistics, and computational linguistics, and identifies challenges and future priorities relating to the following issues: (1) analyzing the distributional characteristics of linguistic constructions and their meanings in large collections of language that are representative of the language that learners experience, (2) the longitudinal analysis of learner language, (3) Natural Language Processing (NLP) analyses of the dimensions of language complexity.

Running Head:
Cognition, Corpora, and Computing

Keywords:
Corpus linguistics, psycholinguistics, cognitive lingusitics, language learning, usage-based approaches

Contact Information:
Nick C. Ellis
Department of Psychology, University of Michigan
3215 E. Hall 1109
530 Church Street
Ann Arbor MI 48109-1043
USA

0. Usage based approaches to Language Learning

Usage-based linguistics explores how we learn language from our experience of language. It is founded upon established findings from four complementary areas of empirical investigation:

(i) Corpus linguistics demonstrates that language usage is pervaded by collocations and phraseological patterns, that every word has its own local grammar, and that particular language forms communicate particular functions: Lexis, syntax, and semantics are inseparable (see Biber & Reppen, 2015; Sinclair, 1991, for reviews).

(ii) Cognitive linguistics shows how language meaning is grounded in our experience and our physical embodiment which represents the world in particular ways. Language consists of many tens of thousands of constructions—form-meaning mappings, conventionalized in the speech community, and entrenched as language knowledge in the learner's mind. Schematic constructions emerge from the conspiracy of memories of particular exemplars that language users have experienced (see Dabrowska & Divjak, 2015; Tomasello, 2003, for reviews).

(iii) The psychology of learning shows that humans have a range of abilities for implicit associative and statistical learning, concept learning and categorization, and explicit declarative learning and analogy-making. These are relevant to the learning of the symbols, sequences, and patterns of language and that imbue our every waking moment (see Rebuschat & Williams, 2012; Sawyer, 2006, for reviews).

(iv) Psycholinguistics shows that our language processing is sensitive to the statistical regularities of language experience at every level of structure (see Ellis, 2002; Traxler & Gernsbacher, 2011, for reviews).

Together, this research shows that "language is never, ever, ever random" (Kilgarriff, 2005). Not in its usage, not in its acquisition, and not in its processing. It follows that theories of language acquisition and processing that ignore the regularities of usage are missing important characteristics of the problem space, characteristics that might have considerable influence on language learning and processing (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015). We should see how the regularities in each of these domains inter-relate.

The usage-based research program necessitates extensive analysis both of the usage from which learners learn, and of learner usage as it develops, both for first language acquisition (Behrens, 2009) and for second language acquisition (Granger, Gilquin, & Meunier, 2015). This program involves considerable data recording, transcription, and analysis, using a variety of corpus and computational techniques, many of them specially devised for learner language. This paper surveys relevant developments across the psychology of learning, first (L1) and second (L2) language acquisition, psycholinguistics, corpus linguistics, and computational linguistics, and identifies challenges and future priorities relating to the following issues:

1. Analyzing the distributional characteristics of linguistic constructions and their meanings in large collections of language that are representative of the language that learners experience.
2. The longitudinal analysis of learner language.
3. NLP analyses of the dimensions of language complexity.

1. Analyzing Usage

Ellis, Römer, and O'Donnell (2016) report analyses of the usage of a range of verb-argument constructions (VACs) including verb locative (VL), verb object locative (VOL), and ditransitive (VOO) forms in the British National Corpus (BNC, 2007) as a large representative sample of English. Using natural language processing techniques, they explored the distributional properties of a sample of VACs as associations of form and function by analyzing their verb selection preferences in the 100 million words of the BNC. In conjunction, they applied network science methods to determine the semantic network structure of the verbs in these constructions. They also measured factors relevant for learning: the type-token frequency distributions of verbs in VACs, as well as prototypicality, semantic cohesion, and polysemy.

These investigations revealed remarkable patterning. VL, VOL, and VOO VACs are (1) Zipfian in their verb type-token constituency in usage, (2) selective in their verb form occupancy, and (3) coherent in their semantics, with a network structure with a Zipfian distribution of degree. Zipf's law states that in human language, the frequency of words decreases as a power function of their rank 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. (Zipf, 1949). Zipfian degree is where there are a small number of prototypical nodes of high betweenness centrality (which allow small-world connectivity of related constructs in radial categories) along with many lesser connected nodes. Ellis et al. (2016) argue that it is this coming-together of the two Zipfian distributions, those for type-token frequency in usage and for degree in the VAC semantic network, that makes VACs coherent. The VAC pattern is centered upon a high token frequency exemplar that is also prototypical in the action-dynamic construal of events to which it relates.

These are interesting proposals, but much more work is needed to test their extent, detail, and replicability. Ellis et al. (2016) identified a range of limitations and challenges for future research.

1.1. Construction Sampling and Description

The first problem concerned the small sample of VACs. The initial study focused upon less than 50 VACs selected as a convenience sample from over 700 patterns in the Verb Grammar Patterns analyzed in the COBUILD project (Francis, Hunston, & Manning, 1996). The full range of patterns requires analysis in order to determine the generalizability of these findings. Furthermore, the COBUILD approach is but one of a variety of approaches to verb classification in cognitive linguistics. Other theories of construction grammar (Croft, 2012; Diessel, 2013; Fillmore, 1988; Goldberg, 1995; Hilpert, 2014; Taylor, 1998; Tomasello & Brooks, 1999; Trousdale & Hoffmann, 2013) present different, complementary descriptions of English verb grammar which could well be explored in similar fashion in order to determine the associative and statistical patterning in usage, acquisition, and processing of constructions so defined.

1.2. Corpora

Like many other researchers, Ellis et al. (2016) used the BNC because, like Everest, it is large and it is there. The BNC is reasonably balanced and was large by the standards of the mid 1990s. It was a huge accomplishment. But standards change and expectations rise. However well it represents a sum of 1990s English language, it is not representative in detail of any one user. There is much interest within corpus linguistics and psycholinguistics in the ways in which language varies by speakers, genre, and register. Spoken language follows quite a different grammar from written language (Biber, 1988; Brazil, 1995; Leech, 2000). Child-directed speech

has its own very special properties. Language varies as a function of the different demands of crafted, considered, and edited written composition vs. fast on-line processing for spoken production and comprehension. Hare, McRae, and Elman (2004), Gahl, Jurafsky, and Roland (2004), and Roland, Dick, and Elman (2007) show how these different conditions affect VACs in language usage. Analyses of nomothetic corpora such as the BNC are only roughly indicative of the input that any one learner might receive.

1.3. Measuring Meaning

Ellis et al. (2016) present some rudimentary analyses and operationalizations of verb semantics in order to investigate prototype effects (Taylor, 2015), semantic cohesion and polysemy, their measurement in corpora of usage, and their effects upon acquisition and processing. They predicted that VACs with coherent semantics would be acquired before those with less coherent semantics, that polysemous VACs will be harder to acquire than monosemous ones, and that more central meanings within polysemous constructions will be acquired before more peripheral ones.

Like many other researchers, they used WordNet (Miller, 2009) for its convenience as a human-categorized and psychologically informed lexical database. WordNet is the result of very many person-years of lexicographic and psycholinguistic effort into categorizing verb semantics into trees based upon similarity. They then used network-science measures such as average clustering, degree assortativity, transitivity, degree centrality, betweenness centrality, and closeness centrality (de Nooy, Mrvar, & Batagelj, 2010; Newman, 2010; Steyvers & Tennenbaum, 2005) to measure the network properties. They also applied a range of network science metrics to identify polysemous communities of meaning as groups of nodes with dense

connections to the other nodes in the group and sparser connections to other nodes posited to belong to different communities, their sizes, their connectedness, etc.

Ellis et al. (2016) demonstrate that verbs that are central in the VAC meaning network are earlier acquired, and they also show effects of verb betweenness-centrality in VAC processing. However, much remains to be done regarding the other questions relating to polysemy and network density on VAC acquisition.

There are many issues relating to the mapping from text to meanings. High frequency words are polysemous and WordNet deals with these by allocating them to different 'synsets'. For example, the lemma THINK occurs in 13 different synsets and KNOW in 11. Without carrying out word sense disambiguation to determine which sense of THINK to compare with which sense of KNOW, Ellis et al. (2016) calculated similarity scores for each of the 143 possible synset pairs and use the maximum value: For THINK and KNOW this value of path similarity is 0.5 resulting from the synset pair "think#v#1" and "know#v#11". It would be better to identify the particular sense that was intended for each verb in its textual context. Methods for word-sense disambiguation are improving all of the time (Agirre & Edmonds, 2007; Jurafsky & Martin, 2009; Kilgarriff, 1998), although they can be computationally expensive. Future availability of sense-tagged corpora is essential if we are to properly explore constructions as form-meaning mappings.

The problems described for analyzing type-token distributions in nomothetic corpora apply equally to the measurement of meaning in nomothetic corpora. Network measures of semantics based upon hundreds of millions of words approximate English usage as a whole. But each of us has our own experience, and it would be good to get somehow closer to idiographic semantics. Furthermore, we are more interested in growth and learning than we are static descriptions, and

so there is need to develop methods for the analysis of longitudinal growth in semantic networks (Steyvers & Tennenbaum, 2005).

Finally, while WordNet is a unique lexico-semantic resource, it provides labeled word meanings and their associations rather than grounded perceptual or motor representations corresponding to the denoted entities and events (Bergen & Chang, 2012). There is much relevant work on perceptual symbol systems (Barsalou, 1999, 2008), on embodied cognition (Clark, 1998; Lakoff & Johnson, 1999; Rosch, Varela, & Thompson, 1991; Shapiro, 2014) and on Embodied Construction Grammar (Bergen & Chang, 2003, 2012). There are promising developments in categorizing lexical concepts through their commonality in brain imaging that would be relevant as well (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Just, Cherkassky, Aryal, & Mitchell, 2010; Mitchell et al., 2008). These are the rich types of meaning that embodied cognition research recognize and with which usage-based approaches should engage.

The notion that meanings can be identified in individual words is increasingly questioned by those pursuing more dynamic perspectives upon human cognition (e.g., Spivey, 2006). Elman (2011) argued against the construct of the mental lexicon, calling for its replacement by more dynamic, contextualized, connectionist models of the interactions of knowledge gathered from socially situated, embodied, usage. Elman's argument, based on his life experience at the forefront of cognitive science, parallel those of Kilgarriff, who, from the vanguard of lexicography and corpus analysis (Kilgarriff, Rychly, Smrz, & Tugwel, 2004), argued against discrete classification of word senses, seeing them instead as a continuous space of meanings largely defined by the contexts in which a word appears:

> The scientific study of language should not include word senses as objects in its ontology. Where word senses have a role to play in a scientific vocabulary, they are to be constructed as abstractions over clusters of word usages... I don't believe in word senses (Kilgarriff, 1997, p. 25).

Kilgarriff's computational lexicographic conclusions clearly echo prior philosophical investigations of the 'language-game' wherein "the meaning of a word is its use in the language" (Wittgenstein, 1953, p. 43), linguistic analyses of the 'context of situation' whereby "you shall know a word by the company it keeps" (Firth, 1957, p. 11), and corpus linguistic recognition of "the phrase, the whole phrase, and nothing but the phrase" as the unit of meaning (Sinclair, 1991, 1996, 2005).

The development of valid models of verb semantics that could be applied at the scale of the current research is perhaps the greatest challenge for cognitive linguistics and usage-based approaches more generally.

2. Acquisition

2.1. L1 Acquisition

Usage-based accounts of language acquisition hold that children learn linguistic constructions from the conspiracy of experienced exemplars, with abstract syntactic constructions and their associated meanings emerging from the statistical distribution of form-function correspondences in usage (Ambridge & Lieven, 2011; Goldberg, 2006; Tomasello, 2003). Associative learning theory suggests that the learning of VACs as categories will be affected by characteristics of type-token frequency distribution (including verb frequency in the VAC, and the degree to which verbs are faithful in their association with particular VACs), along

with factors relating to VAC semantics (including network density and verb prototypicality as network centrality). Ellis et al. (2016) tested these ideas by means of a large-scale analysis of VACs to determine the distribution of forms in all the UK and USA English child-directed speech (CDS; 4,809,299 words) and accompanying child language (2,559,260 words) available in the CHILDES database (MacWhinney, 2000b). These showed (1) the distribution of verb types in a range of schematic as well as more specific VACs is near-Zipfian, (2) VACs are selective in their constituency, and (3) VACs are semantically coherent. Children's VAC acquisition follows these patterns, being affected by input frequency, contingency, and semantic prototypicality. Child VACs are seeded by the more frequent and semantically prototypical verbs that occupy the VAC in the input and these verbs continue ever to lead VAC acquisition. Child frequency of verb usage in particular VACs follows adult verb frequency of usage in these VACs with *r value*s ≈ 0.8.

These are interesting demonstrations of the structure latent in the input and of what children can extract from it, but much more work is needed to test the extent, detail, and replicability of these findings.

2.1.1. Limitations and Future Research Priorities

Firstly, naturalistic corpora comprise the correlated language of speakers in grounded situations -- conversation partners are talking together about matters of shared concern. The language in the CHILDES corpora comes from a variety of situations and tasks, but generally represents everyday, common-ground interactions. Thus overlap between child and adult language might reflect what is being talked about as much as it might reflect shared language competencies. Learner language corpora show what learners say; they do not show what they know. Experimental techniques are needed to probe aspects of knowledge and understanding

(Ambridge & Rowland, 2013). Child language research "has been greatly improved by the development of a range of methods: e.g. preferential looking studies, 'weird word order' studies, and by the use of novel verbs. Priming, eye-tracking, and EEG studies hold out further promise of being able to monitor the types of cue that children use and the scope of their abstractness as these constructions are processed" (Lieven, 2014, p. 52).

The analyses of verb semantics are extremely simple. They suffer from focusing upon words (as considered by adults) rather than embodied mental simulations, or the conscious content of young children. It is a further stretch to try to get a handle on developing infant semantics using WordNet as a first resource.

Corpus searches of learner language are fairly good at finding when a learner has produced the target form, or some close approximation, but NLP taggers and parsers have difficulty dealing with learner language. Thus findings chart early and subsequent success, much better than they do errors on the road to acquisition. Research in error tagging and annotation is clearly a priority (Lüdeling & Hirschmann, 2015).

Ellis et al. (2016) describe broad cross-sectional analyses over very many individual children and parents. Whatever the advantages of large-scale analysis, this is a thin nomothetic stew. There is a pressing need for large-scale individual longitudinal analyses using large, dense corpora (Behrens, 2008; Lieven, Behrens, Speares, & Tomasello, 2003). There should be a focus upon individual children, analyzing at the level of the dyadic exchange, the shared focus of attention, the alignment, the linguistic exchange and uptake. While this is being attempted for L1 acquisition (Roy, 2009; Tomasello & Stahl, 2004), there is nothing yet comparable for second language acquisition (L2A).

In reflecting on 40 years of the *Journal of Child Language*, MacWhinney (2014) described the overarching problem we have in terms of tracking acquisitional patterns that operate across longer timeframes. Cross-sectional data cannot reveal the dynamic aspects of these processes. We need to understand exactly what interactions and social configurations can lead to acquisition. "To study such processes, we must commit ourselves to increasingly ambitious attempts to record the development of individual children or perhaps several children in similar sociolinguistic contexts" (MacWhinney, 2014, p. 130). Cross-linguistic study is equally essential: it is important to build up the store of data on languages other than English, particularly non-European languages (Berman & Slobin, 1994; Crystal, 2014; Slobin, 2014).

The statistical association of factors like frequency, conditional frequency, contingency, and semantic centrality in the input with emerging child language acquisition has a firm foundation (Brown, 1973; MacWhinney, 1987b), and statistical methods for corpus analysis are currently undergoing considerable sophistication and refinement (Baayen, 2008; Gries & Divjak, 2012; Gries & Ellis, 2015). However, the nature of usage entails that many of these variables are highly correlated in the input, and it is difficult to disentangle them. There are current debates about which measures of association are most relevant to acquisition and processing (Evert, 2005; Gries, 2015; Wiechmann, 2008).

Ultimately, we need models of usage and its effects upon acquisition. There are many factors involved, and research to date has tended to look at each variable by variable, hypothesis by hypothesis, one at a time. But they interact. We need theoretical models of learning, development, and emergence that take these factors into account dynamically:

> Children learn language from what they hear as this interacts with their domain-general cognitive learning processes, the current state of their language system, and

> their communicative intentions. For the usage-based approach, the structure of the input and what children can extract from it is crucial. The collection and transcription of large corpora of children's and their caretakers' speech combined with the development of more sophisticated computational techniques has been very important in demonstrating how much information is available in the input and how this might impact on learning. (Lieven, 2014, p. 48)

2.2. L2 Acquisition

There are few available longitudinal corpora of L2A where the sampling is in any depth. The ESL data from the European Science Foundation (ESF) project (Dietrich, Klein, & Noyau, 1995; Feldweg, 1991; Perdue, 1993) is perhaps the richest available. The ESF study, carried out in the 1980s over a period of 5 ½ years, collected the spontaneous second language of adult immigrants in France, Germany, Great Britain, The Netherlands and Sweden. There were in all 5 target second languages (English, German, Dutch, French, and Swedish) and 6 first languages (Punjabi, Italian, Turkish, Arabic, Spanish, Finnish). Data was gathered longitudinally with the learners being recorded in interviews every 4 to 6 weeks for approximately 30 months. The corpus is available from the Max Planck Institute for Psycholinguistics (http://www.mpi.nl/world/tg/lapp/esf/esf.html) and alternatively in CHILDES (MacWhinney, 2000a, 2000b) chat format from the Talkbank website (http://talkbank.org/data/SLA/).

Ellis et al. (2016) analyzed the naturalistic second language acquisition of English VACs in the seven EFL learners in the ESF corpus (Perdue, 1993). As for the L1 analyses described in 2.1, in the naturalistic L2A of English, VAC verb type/token distribution in the input is Zipfian with learners first acquiring the most frequent, prototypical and generic exemplar (e.g. *put* in

VOL, *give* in VOO, etc.). Correlations between learner uptake and frequency of lemma use in the input are in excess of $r > 0.89$. The first-learned verb in each VAC is prototypical of that construction's action semantics but also generic and thus widely applicable. Other verbs which fit the VAC prototype well, but which have additional specifications of manner which restrict their usage, tend to be acquired later. The first-learned verbs in each VAC are distinctively associated with that construction in the input. Correlations between learner uptake and contingency are of the order *r value*s ≈ 0.9. Thus VAC acquisition was affected by the frequency and frequency distribution of exemplars within each slot of the construction (e.g. [Subj V Obj $Obl_{path/loc}$]), by their prototypicality, and by their contingency of form-function mapping. Again, these findings are supportive of usage-based explanations, but again there are many limitations which need to be addressed in further research.

2.2.1. Limitations and Future Research Priorities

The ESF project, as labor-intensive and as groundbreaking as it was, is still far too small to chart longitudinal development, or to catch other than the most frequent VACs. It was designed for a different purpose -- to look at multiple L1/L2 contrasts in naturalistic L2A. We need dense corpora for L2 acquisition. If language learning is in the social cognitive linguistic moment, we need to capture all these moments, so that we can objectively study them (Ortega & Iberri-Shea, 2005). Conversation analysis (commonly abbreviated as CA) is an approach to the study of social interaction, embracing both verbal and non-verbal conduct, in situations of everyday life. We need large dense longitudinal corpora of L2 use, with audio, video, transcriptions and multiple layers of annotation, for data sharing in open archives. We need these in sufficient dense mass that we can chart learners' usage history and their development. We need them in sufficient detail that we can get down to the fine detail of conversation analyses of the moment

(Kasper & Wagner, 2011; Markee, 2008; Markee & Kunitz, 2013). MacWhinney has long been working towards these ends, first with CHILDES (MacWhinney, 2000b), then with Talkbank (MacWhinney, 2007). These projects have developed a number of CLAN tools for computer analyses of large bodies of data. To allow accompanying rich, moment-by-moment description, MacWhinney and Wagner (2010) have also been developing tools for fine grained Conversation Analysis (CABank).

The Interaction Approach to SLA (Gass, 2003; Long, 1996) emphasizes how learners benefit from social interaction because of a variety of developmentally helpful opportunities, conditions, and processes which interaction can expose them to. These include input, negotiation, output, feedback, and attention. Interaction Approach research shows the importance of negotiation, where participants are focused on resolving a communication problem and this "connects input, internal learner capacities, particularly selective attention, and output in productive ways" (Long, 1996, p. 452). Interaction-partners often focus learner attention by means of a clarification request, or negative feedback, or correction, or focus-on-form, or explicit instruction, recruiting consciousness to overcome implicit routines that prove non-optimal for joint understanding (Ellis, 2008b; Gass, 2003; Mackey & Gass, 2006). The opportunities of rich multimodal data afforded by platforms such as CABank allow the study of the cognitive alongside the social (Douglas Fir Group (Atkinson, 2016; Ellis, 2015). Such efforts involve a major commitment, but they make a huge contribution too. For L1 and L2 both, corpora provide "a level playing field on which debates about the import of language sample data, especially longitudinal data, can be played out" (Fletcher, 2014, p. 18).

Language learning is a sampling problem. Learners have to estimate the system from a limited sample (Ellis, 2008a). If we want to understand their acquisition, then we need

representative samples of their usage history. Zipfian distributions mean that as researchers we have a better handle on the most frequent constructions. But there is considerable variability at lower frequencies (Tomasello & Stahl, 2004). Learning environments vary tremendously too. A course textbook might serve as a better guide for foreign language learners. Classroom discourse is highly variable and deserves corpus investigations in its own right, as persuasively demonstrated by Collins, Trofimovich, White, Cardoso, and Horst (2009) (see also Collins & Ellis, 2009), as does analyzing the language used in foreign language teaching textbooks (Biber, 2006; Gouverneur, 2008; Römer, 2004).

The last twenty years have seen impressive developments in Learner Corpora (Granger et al., 2015). Notable achievements beyond Talkbank include the International Corpus of Learner English (ICLE; Granger, Dagneaux, Meunier, & Paquot, 2009), the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin, De Cock, & Granger, 2010), from the Education First (EF) research unit at the University of Cambridge UK, the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen, Alexopoulou, & Korhonen, 2013), and the many contributions from University Centre for Computer Corpus Research on Language (UCREL) and the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University.

A significant issue with searching learner language is that one can find instances where the learner has produced the target form, but it is much harder to find instances where they have produced some non-targetlike variant. Dealing with learner language with NLP techniques poses considerable challenges, although significant developments are being made (Meurers, 2015), likewise there are important developments in human annotation of error (Dickinson & Lee 2009; Lüdeling & Hirschmann, 2015).

If we are serious in our investigations of usage-based linguistics, in the charting of learner language as it develops in communication, ultimately, we need both nomothetic and dense idiographic longitudinal corpora. This will involve invested interactions between corpus linguists, cognitive linguists, and researchers of second language acquisition. We need intensive efforts to build up the store of L1 and L2 data in English, in a wide range of languages other than English, and particularly in non-European languages.

3. Measuring the Complexity of Language

Achieving a coherent, comprehensive, and valid conceptualization of language complexity is an essential foundation for research in both L1 and L2 learning. It would permit a variety of research goals including those relating to *acquisition*: 'Does interlanguage complexity increase as second language users become more proficient in the target language?'; to *instruction and task design*: 'Does interlanguage complexity systematically vary (i.e., can we see reliably lower, same, or higher complexity) depending on the cognitive task conditions imposed on L2 users?'; and to *measurement and testing*, benchmarking developmental level: 'Does the interlanguage complexity of production increase with growing grammatical development?' (Ortega, 2012).

There are a number of different dimensions of complexity which all require operationalization before we can answer these important questions in their regard. These are (1) structural complexity: morphological, lexical, and syntactic dimensions, (2) phraseology, collocation, and lexico-grammar, and (3) statistical measures of proficiency

3.1. Structural complexity: Morphological, Lexical, and Syntactic Dimensions

Historically, the most widely employed operationalizations of structural complexity have relied heavily on the notions of length (e.g., average number of words per T-unit) and density of subordination (e.g., average number of finite clauses per T-unit). But there is a range of possibilities, and in their important review of 40 studies of L2 structural complexity over the years 1995-2008, Bulté and Housen (2012) demonstrate the use of no less than 40 different operationalizations, with 22 studies using one or two measures only (see Table 2, p. 32). They conclude "the link between theoretical characterizations of complexity and the way in which complexity has been operationalized in CAF [Complexity, Accuracy and Fluency] research has not been explicit enough" (p. 42).

Pallotti (2015) provides a considered and constructive overview of linguistic definitions of textual structural complexity in terms of the number of different elements and their interconnections (i.e. their systematic, organized relationships). Once these are determined, a measure of complexity can be expressed in terms of the length of the shortest description that is needed to represent them (e.g., Kolmogorov complexity, see Ehret & Szmrecsanyi, 2015). He regards linguistic complexity as an absolute, an objective, inherent complexity that is clearly quantifiable in texts. His clear and constant goal is that of "a simple, coherent view of the construct… defined in a purely structural way," explicitly excluding cognitive cost (difficulty) and developmental dynamics (acquisition) from this theoretical definition and its operationalization (p. 177).

Pallotti considers elements at morphological, syntactic, and lexical levels. Assessing a text's morphological complexity involves counting, "for each word class (nouns, verbs,

adjectives, etc.), its exponents, i.e. the forms taken by lexemes to express grammatical categories and functions" (p. 121). He defines syntactic complexity as "the number of interconnected constituents in a structure, which is the principle behind three measures such as length of phrase, number of phrases per clause and number of clauses per unit." (p. 123). He defines lexical complexity in terms of lexical diversity (Jarvis, 2013) which "can be gauged basically by looking at type/token ratios, with subsequent refinements proposed to overcome the effects of text length, such as the Guiraud index and D" (p. 125). He concludes that these various measures could be applied individually, as one might be interested in studying a certain type of complexity only, e.g. lexical or syntactic, or they might be used together, to provide a global estimate of a text's complexity, and he outlines steps by which such combination could be achieved.

Given that operationalizations of objective text-based complexity can be fairly easily applied to texts and corpora using tools that have become readily available (such as, for morphological diversity, the morphological complexity tool (Brezina & Pallotti, 2015); for various measures of syntactic complexity the programs developed by Lu (2010; 2011); for lexical diversity (*vocd*, Malvern, Richards, Chipere, & Duran, 2004); and generally the suite of programs developed over many years and freely provided in CLAN (MacWhinney, Fromm, Forbes, & Holland, 2011), they will be put to much use over the next few years.

While these operationalizations provide means to separately measure morphological, syntactic, and lexical complexity in these ways, research and theory in language learning, cognitive linguistics, and other usage-based approaches strongly counter assumptions of modularity and the orthogonality of morphological, syntactic, and lexical dimensions of language. Nevertheless, these are the traditional units and divisions of linguistic theory, and it is therefore sensible to assess the degree of their interplay.

These operationalizations stemming from linguistic theory need supplementing with measures informed by theories of acquisition and psycholinguistic processing because they miss important aspects of language complexity relating to the statistical learning of language and collocations, chunks, and other emergent constructions which relate lexis, syntax and semantics.

3.2. Phraseology, Collocation, and Lexico-grammar

Usage-based linguistics holds that much of language is based on memorized chunks. Corpus-linguistic analyses show that the *Principle of Idiom* pervades usage and language knowledge: "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (Sinclair, 1991: 100).

Both the type-token ratio and the lexical diversity score for *a lot of words* and *a rich, diverse vocabulary* are the same, though the latter demonstrates a greater potential command of English lexis. Vocabulary testers have long taken word frequency into account in their measurement of vocabulary depth because lower frequency words are more telling (Cobb, 2010, 2015; Heatley, Nation, & Coxhead, 2002; Nation, 2001).

Frequency and probability are equally essential to syntactic complexity. The noun phrase complexity scores for *blazing row*, *heated dispute, ?heated row* and ?*blazing dispute* are all the same, as are those for *very attached*, *very affected*, *very punished*, and *closely attached*, *deeply affected*, and *severely punished*. Yet you recognize that some of these pairs of words go together well, while others clash, despite their being equally good syntactically. It is now firmly established that language users have rich knowledge of such collocations (for reviews see Ellis, 2002, 2012). Crossley, Salsbury, and Mcnamara (2014) analyzed a corpus of 240 spoken texts

and 240 written texts produced by beginning L2 learners, intermediate L2 learners, advanced L2 learners, and native speakers which had been scored for both analytic and holistic features of lexical proficiency by trained raters. Using a multiple regression analysis, the study found that while collocation accuracy, lexical diversity, and word frequency were all significant predictors of human evaluations of lexical proficiency, collocation accuracy explained the greatest amount of variance in the holistic scores (84% in the written samples and 89% in the spoken samples).

Statistical syntagmatic knowledge does not stop at bigrams: it pervades phraseology and formulaic language (Schmitt, 2004). We can measure these complex patterns in corpora, and we can separately show that users have knowledge of them (Gilquin & Gries, 2009; Gries & Divjak, 2012; Gries & Ellis, 2015; O'Donnell, Römer, & Ellis, 2013). In developmental dynamics, this knowledge differentiates first and second language speakers: L2 learners typically do not achieve nativelike idiomaticity (Granger, 2001; Pawley & Syder, 1983).

Empirical demonstration of the co-selection of grammar and lexis and their importance in the assessment of structural complexity comes from analysis of 'criterial features', the textual characteristics which differentiate learner essays awarded different grades on the Common European Framework of Reference (CEFR) levels. The English Profile Programme, a series of studies examining the language produced at each of the CEFR levels, presents empirical findings on lexico-grammatical features and functional progression of English in the Cambridge Learner Corpus of written English scripts from the Cambridge ESOL examinations, covering the proficiency range from A2 to C2, and containing around 45 million words (as of 2011) (Hawkins & Buttery, 2010; Hawkins & Filipović, 2012). Sample grammatical Criterial Features for B2 Level (UCLES /CUP, 2011, pp. 21-22) include:

B2.6	The verbs *appear, cease, fail, happen, prove, turn out*, and the adjectives *certain, likely, sure, unlikely* + infinitive [Subject-to-Subject Raising, NP-V-VPinfin]	To my regret, the evening totally *failed to live up* to expectations.
B2.7	*imagine, prefer* + object + infinitive [Subject-to-Object Raising, NP-V-NP-VPinfin]	I would *prefer* my accommodation to be in log cabins…
B2.8	the verbs *expected, known, obliged, thought* (in Passive voice) + infinitive [Subject-to-Object Raising plus Passive, NP-V-NP-VPinfin]	Your theatre *is known to present* excellent spectacles.

Every word has its own local grammar (Hunston & Francis, 2000; Sinclair, 1991, 1996). Lexis and grammar co-select. There is no sense then to measure them as separate independent measures of syntactic and lexical complexity.

Corpus-derived metrics can now readily be applied to texts. Some relevant techniques have been available in Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) for several years. Recently, however, more sophisticated and more usable software has become freely available, and more is sure to follow. Vajjala and Meurers (2012) illustrate the potential of combined lexical and syntactic indices of text complexity in NLP classification, developments that were inspired by SLA research. Kyle and Crossley (2014) describe the Tool for the Automatic Analysis of LExical Sophistication (TAALES), which calculates text scores for 135 classic and newly developed lexical indices related to word frequency, range, bigram and trigram

frequency, academic language, and psycholinguistic word information. Kyle and Crossley showed that TAALES indices explained 48% of the variance in holistic ratings of second language (L2) learner lexical proficiency and 49% of the variance in ratings of L2 speaking proficiency. The strongest predictor of speaking proficiency was trigram written frequency, which accounted for 35% of the variance in holistic speaking proficiency scores: speaking samples that had more high-frequency trigrams tended to receive higher scores. They conclude that "although frequency may indeed be an important indicator of written lexical proficiency and spoken proficiency, range and n-gram indices may be even more important" (pp. 773-774).

3.3. Statistical measures of proficiency

Psycholinguistic research into cognitive complexity finds that statistical knowledge of co-occurrences pervades language processing. It underpins fluent sentence processing (see MacDonald & Seidenberg, 2006, for review). Eye-movement research shows that the fixation time on each word in reading is a function of the frequency of that word (frequent words have shorter fixations) and of the forward transitional probability (the conditional probability of a word given the previous word $P(w_k|w_{k-1})$: for example, the probability of the word *in* given that the previous word was *interested* is higher than the probability of *in* if the last word was *dog*) (McDonald & Shillcock, 2003, 2004). Generally, analyses of large corpora of eye-movements recorded when people read text demonstrate that measures of surprisal account for the costs in reading time that result when the current word is not predicted by the preceding context (Demberg & Keller, 2008). The surprisal of a word in a sentential context corresponds to the probability mass of the analyses that are not consistent with the new word. Surprisal is inversely related to probability. Research operationalizations of surprisal in language involves computing norms in corpora of usage, and then looking for violations of those norms. The simplest possible

case is the unconditional probability (i.e., relative frequency) of, say, a word in a corpus. '*The…'* is less surprising than is '*Discombobulate…'*. A slightly more complex example is a simple forward transitional probability such as the probability of the word *y* directly following the word *x* (compare *'strong tea', 'strong computers', 'powerful tea', 'powerful computers'*), or a conditional probability such as the probability of a particular verb given a construction ('*give'* is much more likely in a ditransitive than is '*kick'*) (Gries & Ellis, 2015; Gries & Stefanowitsch, 2004). Measuring surprisal requires a probabilistic notion of linguistic structure (utilizing transitional probabilities or probabilistic grammars).

Consider too the statistics of linguistic construction *qua* symbol. Linguistic structures convey meaning. Constructions, from fixed words and expressions to abstract morphology and syntax, are symbols, mappings of linguistic form and their interpretation. They do so with differing degrees of transparency and reliability. Psychological research into associative learning has long recognized that while frequency of form is important, more so is *contingency*. Cues with multiple interpretations are ambiguous and so hard to resolve, whereas cue-outcome associations of high contingency are reliable and readily processed. Contingency of cue-outcome mapping is a driving force of all associative learning, and is central in the Competition Model (Bates & MacWhinney, 1987) and other models of the rational learning and processing of form-function constructions.

Every sentence, every part of text, contains many such cues to interpretation, and the interpretation of the sentence as a whole involves constraint-satisfaction: the conspiracy of, and competition between, these cues (Bates & MacWhinney, 1987; MacWhinney, 1987a). To understand these aspects of cognitive complexity, we need to measure the validities and reliabilities of these mappings in corpora (Kempe & MacWhinney, 1998). The reliability or

contingency of the association between the linguistic structure and its semantics likewise determines the learnability of linguistic symbols, as it does every other association (Ellis, 2006). Every linguistic form is ambiguous. In acquisition, learners have to figure stimuli out: to learn the probability distribution P(*interpretation*|*cue, context*), the probability of an interpretation given a stimulus cue. Usage-based linguistics believe that this figuring is achieved, and cognition optimized, by the implicit tallying of the frequency, recency, and context of linguistic structures in the learner's first person, intentional, conscious, contextualized, and embodied experience (Yurovsky, Smith, & Yu, 2013). We need to measure these aspects of usage in rich longitudinal learner corpora.

4. Conclusions

Usage-based approaches hold that syntactic structure, lexis, and semantics are inseparable, and that they go together probabilistically. The very essence of language comes from their inter-relations in usage. However, we have a long research program ahead of us to operationalize these different aspects of complexity and the degree of their associations in order both to chart learning and development and to inform language assessment. This program has to involve collaborations between researchers of cognition, corpora, and computing, as well as more social/socio-cultural/interactionist researchers too, because language learning emerges from everyday usage experience, with its attendant dynamics of cognition, attention, consciousness, social interaction, cultural scaffolding, zone-of-proximal development, and phraseological old Uncle Tom Cobley and all.

6544 words

5. References

Agirre, E., & Edmonds, P. (Eds.). (2007). *Word sense disambiguation: Algorithms and applications*. New York: Springer.

Ambridge, B., & Lieven, E. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge: Cambridge University Press.

Ambridge, B., & Rowland, C. F. (2013). Experimental methods in studying child language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(2), 149-168. doi: 10.1002/wcs.1215

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660. doi: 10.1017/s0140525x99002149

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645. doi: 10.1146/annurev.psych.59.103006.093639

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157-193). Hillsdale, NJ: Lawrence Erlbaum Associates.

Behrens, H. (2008). *Corpora in language acquisition research: History, methods, perspectives*. Amsterdam: John Benjamins.

Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. *Linguistics, 47*, 383-411. doi: 10.1515/ling.2009.014

Bergen, B., & Chang, N. C. (2003). Embodied construction grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction grammars:*

Cognitive grounding and theoretical extensions (pp. 147-190). Amsterdam/Philadelphia: John Benjamins.

Bergen, B., & Chang, N. C. (2012). Embodied construction grammar. In G. Trousdale & T. Hoffmann (Eds.), *Oxford handbook of construction grammar* (pp. 168-190). Oxford: Oxford University Press.

Berman, R. A., & Slobin, D. I. (Eds.). (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, N.J.: Lawrence Erlbaum.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University language*. Amsterdam: John Benjamins.

Biber, D., & Reppen, D. (Eds.). (2015). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

BNC. (2007). BNC XML Edition http://www.natcorp.ox.ac.uk/corpus/.

Brazil, D. (1995). *A grammar of speech.* Oxford: Oxford University Press.

Brezina, V., & Pallotti, G. (2015). Morphological complexity tool, available from http://corpora.lancs.ac.uk/vocab/analyse_morph.php.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam/Philadelphia: Benjamins.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. Cambridge, MA: The MIT Press.

Cobb, T. (2010). Learning about Language and Learners from Computer Programs. *Reading in a Foreign Language, 22*(1), 181-200.

Cobb, T. (2015). The Compleat Lexical Tutor (v.7.0). http://www.lextutor.ca/. http://132.208.224.131/

Collins, L., & Ellis, N. C. (2009). Input and second language construction learning: Frequency, form, and function. *Modern Language Journal, 93*(2), Whole issue. doi: 10.1111/j.1540-4781.2009.00893.x

Collins, L., Trofimovich, P., White, J., Cardoso, W., & Horst, M. (2009). Some input on the easy/difficult grammar question: An empirical study. *The Modern Language Journal, 93*(3), 336-353. doi: 10.1111/j.1540-4781.2009.00894.x

Croft, W. (2012). *Verbs: Aspect and causal structure*. Oxford: Oxford University Press.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2014). Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy. *Applied Linguistics*, 1-22. doi: 10.1093/applin/amt056

Crystal, D. (2014). Editorial. *Journal of Child Language, 41 S1*, v-vi. doi: 10.1017/s0305000914000129

Dabrowska, E., & Divjak, D. (Eds.). (2015). *Handbook of cognitive linguistics*. Berlin: Mouton DeGruyter.

de Nooy, W., Mrvar, A., & Batagelj, V. (2010). *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*, 193-210. doi: 10.1016/j.cognition.2008.07.008

Dickinson, M., & Lee , C. M. (2009). Modifying Corpus Annotation to Support the Analysis of Learner Language. *CALICO Journal, 26*(3).

Diessel, H. (2013). Construction grammar and first language acquisition. In G. Trousdale & T. Hoffmann (Eds.), *The Oxford handbook of construction grammar* (pp. 347-364). Oxford: Oxford University Press.

Dietrich, R., Klein, W., & Noyau, C. (Eds.). (1995). *The acquisition of temporality in a second language*. Amsterdam: John Benjamins.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences, 19*(10), 603-615. doi: 10.1016/j.tics.2015.07.013

Douglas Fir Group (Atkinson, D., Byrnes, H., Doran, M., Duff, P., Ellis, N., Hall, J. K., Johnson, K., Lantolf, J., Larsen–Freeman, D., Negueruela, E., Norton, B., Ortega, L., Schumann, J., Swain, M., and Tarone, E.). (2016). A transdisciplinary framework for SLA in a multilingual world. *Modern Language Journal, 100*, 19-47. doi: 10.1111/modl.12301

Ehret, K., & Szmrecsanyi, B. (2015). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity and Isolation.* Berlin: de Gruyter.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143-188. doi: 10.1017/s0272263102002024

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27*(1), 1-24. doi: 10.1093/applin/ami038

Ellis, N. C. (2008a). Optimizing the input: Frequency and sampling in usage-based and form-focussed learning. In M. H. Long & C. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 139-158). Oxford: Blackwell.

Ellis, N. C. (2008b). The psycholinguistics of the interaction hypothesis. In A. Mackey & C. Polio (Eds.), *Multiple perspectives on interaction in SLA: Second language research in honor of Susan M. Gass* (pp. 11-40). New York: Routledge.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17-44. doi: 10.1017/s0267190512000025

Ellis, N. C. (2015). Cognitive and social aspects of learning from usage. In T. Cadierno & S. W. Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 49-73). Berlin: DeGruyter Mouton.

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Language usage, acquisition, and processing: Cognitive and corpus investigations of construction grammar*. Malden, MA: Wiley-Blackwell.

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon, 6*, 1-33. doi: 10.1075/ml.6.1.01elm

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations.* University of Stuttgart, Stuttgart.

Feldweg, H. (1991). *The European Science Foundation Second Language Database*. Nijmegen: Max-Planck-Institute for Psycholinguistics.

Fillmore, C. (1988). The mechanisms of construction grammar. *Berkeley Linguistics Society, 14*, 35–55. doi: 10.3765/bls.v14i0.1794

Firth, J. R. (1957). *Papers in linguistics: 1934-1951*. London: Oxford University Press.

Fletcher, P. (2014). Data and beyond. *Journal of Child Language 41 S1*, 18-25. doi: 10.1017/s0305000914000191

Gahl, S., Jurafsky, D., & Roland, D. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods, Instruments, and Computers, 36*, 432-443. doi: 10.3758/bf03195591

Gass, S. (2003). Input and interaction. In C. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 224-255). Oxford: Blackwell Publishers

Geertzen, T., Alexopoulou, T., & Korhonen, A. (2013). *Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT).* Paper presented at the 31st Second Language Research Forum (SLRF), Carnegie Mellon University.

Gilquin, G., De Cock, S., & Granger, S. (2010). *LINDSEI: Louvain international database of spoken English interlanguage*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Gilquin, G., & Gries, S. Th. (2009). Corpora and experimental methods: a state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*, 1-26. doi: 10.1515/cllt.2009.001

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gouverneur, C. (2008). The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-driven approach to EFL textbook analysis. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 223-243). Amsterdam: John Benjamins.

Granger, S. (2001). Prefabricated patterns in Advanced EFL writing: Collocations and folmulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *ICLE: International corpus of learner English*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

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

Gries, S Th. (2015). More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff (2013). *Cognitive Linguistics, 26(3)*, 505-536. doi: 10.1515/cog-2014-0092

Gries, S. Th., & Divjak, D. S. (Eds.). (2012). *Frequency effects in language learning and processing*. Berlin & New York: Mouton de Gruyter.

Gries, S. Th., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Currents in Language Learning, 2*, 228-255. doi: 10.1111/lang.12119

Gries, S. Th., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics, 9*, 97-129. doi: 10.1075/ijcl.9.1.06gri

Hare, M., McRae, K., & Elman, J. L. (2004). Admitting that admitting verb sense into corpus analyses makes sense. *Language and Cognitive Processes, 19*, 181-224. doi: 10.1080/01690960344000152

Hawkins, J. A., & Buttery, P. (2010). Criterial Features in learner corpora: Theory and illustrations. *English Profile Journal, 1*, 1-23. doi: 10.1017/s2041536210000103

Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework* (Vol. 1): Cambridge University Press.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx.

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453-458. doi: 10.1038/nature17637

http://www.nature.com/nature/journal/v532/n7600/abs/nature17637.html - supplementary-information

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87-106. doi: 10.1111/j.1467-9922.2012.00739.x

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Second Edition.* Englewood Cliffs, NJ: Prentice-Hall.

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE, 5*, e8622. doi: 10.1371/journal.pone.0008622

Kasper, G., & Wagner, J. (2011). A conversation-analytic approach to second language acquisition. Alternative approaches to second language acquisition. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 117-142.). Abingdon: Routledge.

Kempe, V., & MacWhinney, B. (1998). The acquisition of case-marking by adult learners of Russian and German. *Studies in Second Language Acquisition, 20*, 543-587. doi: 10.1017/s0272263198004045

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*(2), 91-113. doi: 10.1023/a:1000583911091

Kilgarriff, A. (1998). Senseval: An exercise in evaluating word sense disambiguation programs. *Proceedings of the first international conference on language resources and evaluation, Granada, Spain*, 581-588.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory, 1*(2), 263-276. doi: 10.1515/cllt.2005.1.2.263

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwel, D. (2004). *The Sketch Engine.* Paper presented at the EURALEX 2004 Lorient, France.

Kyle, K., & Crossley, S. A. (2014). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*. doi: 10.1002/tesq.194

Lakoff, G., & Johnson, M. H. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. London: Harper Collins.

Leech, L. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*, 675-724. doi: 10.1111/0023-8333.00143

Lieven, E. V. M. (2014). First language development: A usage-based perspective on past and current research. *Journal of Child Language 41*(S1), 48-63. doi: 10.1017/s0305000914000282

Lieven, E. V. M., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage based approach. *Journal of Child Language, 30*, 333-370. doi: 10.1017/s0305000903005592

Long, M. H. (1996). The role of linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisitiom* (pp. 413-468). San Diego: Academic.

Lu, X. (2010). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics, 14*(1), 3-28. doi: 10.1075/ijcl.14.1.02lu

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly, 45*(1), 36-62. doi: 10.5054/tq.2011.240859

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135-157). Cambridge: Cambridge University Press.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd Edition ed., pp. 581-611). London: Elsevier Inc.

Mackey, A., & Gass, S. (2006). Pushing the methodological boundaries in interaction research: An Introduction to the special issue. *Studies in Second Language Acquisition, 28*, 169-178. doi: 10.1017/s0272263106060086

MacWhinney, B. (1987a). The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249-308). Hillsdale, NJ: Erlbaum.

MacWhinney, B. (2000a). *The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs* (3rd ed.). Mahwah NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2000b). *The CHILDES Project: Tools for analyzing talk, Vol 2: The database (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2007). *The TalkBank Project* Retrieved from http://repository.cmu.edu/cgi/viewcontent.cgi?article=1175&context=psychology

MacWhinney, B. (2014). What have we learned? *Journal of Child Language, 41*(S1), 124-131. doi: 10.1017/s0305000914000142

MacWhinney, B. (Ed.). (1987b). *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*, 1286-1307. doi: 10.1080/02687038.2011.589893

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gespraechsforschung, 11*, 154-173.

Malvern, D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke,, UK: Palgrave Macmillan.

Markee, N. (2008). Toward a learning behavior tracking methodology for CA-for-SLA. *Applied Linguistics, 29*, 404-427. doi: 10.1093/applin/amm052

Markee, N., & Kunitz, S. (2013). Doing planning and task performance in second language acquisition: An ethnomethodological respecification. *Language Learning, 63*, 629-664. doi: 10.1111/lang.12019

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, M.A.: Cambridge University Press.

Meurers, D. (2015). Learner language and Natural Language Processing. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 537-566). Cambridge: Cambridge University Press.

Miller, G. A. (2009). WordNet - About us. Retrieved March 1, 2010, from Princeton University http://wordnet.princeton.edu

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*, 1191-1195. doi: 10.1126/science.1152876

Myles, F. (2004). From Data to Theory: the Over-Representation of Linguistic Knowledge in SLA. *Transactions of the Philological Society, 102*(2), 139-168. doi: 10.1111/j.0079-1636.2004.00133.x

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.

O'Donnell, M. B., Römer, U., & 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. doi: 10.1075/ijcl.18.1.07odo

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Szmrecsanyi & B. Kortmann (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127-155). Berlin: de Gruyter.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25*, 26-45. doi: 10.1017/s0267190505000024

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117-134. doi: 10.1177/0267658314536435

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-225). London: Longman.

Perdue, C. (Ed.). (1993). *Adult language acquisition: Crosslinguistic perspectives*. Cambridge: Cambridge University Press.

Peters, A. M. (1983). *The units of language acquisition*. New York: Cambridge University Press.

Rebuschat, P., & Williams, J. N. (Eds.). (2012). *Statistical learning and language acquisition*. Berlin: Mouton de Gruyter.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language, 57*(3), 348-379. doi: 10.1016/j.jml.2007.03.002

Römer, U. (2004). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini & D. Stewart (Eds.), *Corpora and language learners* (pp. 151-168). Amsterdam: John Benjamins.

Rosch, E., Varela, F., & Thompson, E. (1991). *The embodied mind*. Boston, MA: MIT Press.

Roy, D. (2009). New horizons in the study of child language acquisition. *Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, 10*.

Sawyer, R. K. (Ed.). (2006). *The Cambridge Handbook of the Learning Sciences*. Cambridge: Cambridge University Press.

Schmitt, N. (Ed.). (2004). *Formulaic sequences*. Amsterdam: Benjamins.

Shapiro, L. (Ed.). (2014). *The Routledge handbook of embodied cognition*. London: Routledge.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (1996). The search for units of meaning. *Textus, IX*, 75-106. doi: 10.1093/applin/amn052

Sinclair, J. M. (2005). The phrase, the whole phrase, and nothing but the phrase. *Paper presented at Phraseology 2005, Louvain*.

Slobin, D. I. (2014). Before the beginning: The development of tools of the trade. *Journal of Child Language, 41*(S1), 1-17. doi: 10.1017/s0305000914000166

Spivey, M. (2006). *The continuity of mind*. Oxford: Oxford University Press.

Steyvers, M., & Tennenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41-78. doi: 10.1207/s15516709cog2901_3

Taylor, J. R. (1998). Syntactic constructions as prototype categories. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (pp. 177-202). Mahwah, NJ: Erlbaum.

Taylor, J. R. (2015). Prototype effects in grammar. In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 562-579). Berlin: DeGruyter Mouton.

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

Tomasello, M., & Brooks, P. (1999). Early syntactic development: A Construction Grammar approach. In M. Barrett (Ed.), *The development of language* (pp. 116-190). London: University College London Press.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language, 31*, 101-121. doi: 10.1017/s0305000903005944

Traxler, M., & Gernsbacher, M. A. (Eds.). (2011). *Handbook of psycholinguistics.* New York: Academic Press.

Trousdale, G., & Hoffmann, T. (Eds.). (2013). *Oxford handbook of construction grammar*. Oxford: Oxford University Press.

UCLES /CUP. (2011). *English Profile. Introducing the CEFR for English (v.1.1)*. Cambridge: Cambridge University Press.

Vajjala, S., & Meurers, D. (2012). *On improving the accuracy of readability classification using insights from second language acquisition.* Paper presented at the Proceedings of the Seventh Workshop on Building Educational Applications Using NLP.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory, 4*(2), 253-290. doi: 10.1515/cllt.2008.011

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Oxford: Blackwell.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science, 16*, 959-966. doi: 10.1111/desc.12036

Zipf, G. K. (1949). *Human behaviour and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.