3

Does Language Zipf Right Along?

Investigating Robustness in the Latent Structure of Usage and Acquisition

NICK C. ELLIS
University of Michigan

MATTHEW BROOK O'DONNELL
University of Michigan

UTE RÖMER
Georgia State University

EACH OF US AS language learners has had different language experiences, but somehow, we have converged on the same general language system. From diverse and often noisy samples, we end up with similar linguistic competence. How so? Do language form, language meaning, and language usage come together across scales to promote robust induction by means of statistical learning over limited samples? The research described here outlines an approach to this question with regard to English verb-argument constructions (VACs), their grammatical form, semantics, lexical constituency, and distribution patterns. Measurement and analysis of large corpora of language usage identifies Zipfian scale-free patterns in VAC type-token frequency and in the structure of their semantic networks. Using methods from cognitive linguistics, corpus linguistics, learning theory, complex systems, and network science, we explore how these latent structures of usage might promote the emergence of linguistic constructions in first and second language acquisition.

Literature Review

We seek an understanding of robust language acquisition. As a child, you engaged your parents and friends by talking about shared interests and using words and phrases that came to mind; this is how you learned language. None of the authors of this paper was privy to this system, but somehow, we have all converged on a similar-enough English to be able to communicate. How did this happen?

We take a two-pronged approach to this question. First, we consider the psychology of learning as applied to linguistic constructions. This is generally the approach to language acquisition as pursued within usage-based linguistics, cognitive linguistics, construction grammar, child language research, second language acquisition

(SLA), and psycholinguistics (Bybee 2010; Goldberg 2006; Robinson and Ellis 2008; Tomasello 2003). These views share the assumption that language acquisition is similar to the rest of cognition in that we learn language like we learn anything else. Of course, human cognition is not simple; there is much to the psychology of learning. The problem-space of language—mapping thoughts to serial sequences of sound—is particularly special, but it is parsimonious to assume that language is subject to the same learning mechanisms and cognitive constraints as the rest of our experience.

Second, for the factors that promote robust acquisition, we look to research within emergentism, dynamic systems theory, and complex adaptive systems (CAS). CAS are characterized by their robustness to different kinds of perturbations, by their scale-free properties, and by their structures emerging from the interactions of agents and components at many levels (Holland 1995; Page 2009), as shown by the recent explorations of *Language as a Complex Adaptive System* (Beckner et al. 2009; Ellis 1998; Ellis and Larsen-Freeman 2006, 2009b; Larsen-Freeman 1997; MacWhinney 1999; Solé et al. 2005).

A significant discovery in the early cognitive analysis of language involved how basic objects in natural categories underpinned the robust acquisition of nouns. Rosch et al. (1976) showed how basic categories—those that carry the most information in clustering the things of the world—are those whose members possess significant numbers of attributes in common, are visually imageable with similar shapes, and have similar associated motor programs. Basic natural categories are organized around prototypes. These prototype exemplars are most typical of the category, similar to many other category members, and not similar to members of other categories. People categorize prototype exemplars (like *robin* as *bird*) faster than those with less common features or feature combinations like *geese* or *penguins* (Rosch and Mervis 1975; Rosch et al. 1976). Basic categories are also those that are the most codable (faster naming), most coded, and most necessary in language (highly frequent in usage). Children acquire basic-category terms like *dog, bird, hammer,* and *apple* earlier than they do their superordinates *animal, tool,* and *fruit*, or subordinates *collie, wren, ball-peen hammer,* and *Granny Smith*. It is reliable visual and motor perceptual experience along with frequent and highly contingent labeling that makes these nouns reliably and robustly learnable, despite individual children experiencing different types of dogs and birds.

Cognitive linguistics, particularly construction grammar, has since extended these ideas to language as a whole. Nouns typically relate to the things of the world, but because language has emerged to describe our experiences, whole sentences are used to describe the doings of nouns. Linguistic constructions that correspond to basic sentence types encode as their prototypical senses events that are basic to human experience—those of something moving, something being in a state, someone causing something, someone possessing something, something causing a change of state or location, someone causing a change of possession, something undergoing a change of state or location, something having an effect on someone, and so on (Croft 2001, 2012; Goldberg 1995; Levin 1993).

Corpus and cognitive linguistics have shown that grammar and semantics are reliably associated, and, in turn, that grammatical patterns and their corresponding events jointly select particular lexical items. Syntax, lexis, and semantics are inextricably intertwined (Sinclair 2004). The meaning of the words of a language and how they can be used in combination depends on our perception and categorization of the world around us. Since we constantly observe and play an active role in this world, we know a great deal about it; this experience and familiarity is reflected in the nature of language. The differing degrees of salience, as well as the prominence of elements involved in situations that we wish to describe, will affect the selection of subjects, objects, adverbials, and other clause arrangements. Figure/ground segregation and perspective-taking—processes of vision and attention—are mirrored in language and have systematic relations with syntactic structure. In language production, what we express reflects which parts of an event attract our attention; depending on how we direct our attention, we can select and highlight different aspects of the frame, thus arriving at different linguistic expressions. The prominence of particular aspects of the scene and the perspective of the internal observer (i.e., the attentional focus of the speaker and the intended attentional focus of the listener) are key elements in determining regularities of association between elements of visuospatial experience and elements of phonological form. In language comprehension, abstract linguistic constructions (like locatives, datives, and passives) guide the listener's attention to a particular perspective on a scene while backgrounding other aspects (Langacker 1987; Talmy 2000; Taylor 2002).

By processes of syntactic and semantic bootstrapping, these associations of form and function could allow linguistic constructions to be learnable in an exemplar-by-exemplar fashion, with abstract schematic patterns induced from particular usage patterns and their interpretations. Researching this possibility requires interdisciplinary collaborations. The investigation of form involves structuralist, corpus linguistic, and computational linguistic approaches; the investigation of function involves functionalist, cognitive linguistic, and psycholinguistic analyses; and the study of embodied force dynamics involves an understanding of semantic organization and more. Their association requires quantitative linguistics for the statistical tallying of form and function as well as an understanding of the psychology of learning. The result of these collaborations should not be a dictionary, a grammar manual, or a frequency list. Rather, it should be a systemic network integrating the syntactic constructions of a language, the lexis they select, their meanings, and the distributions and mappings of these forms and functions.

In what follows, we sketch how we believe this work might progress, illustrating it with some preliminary findings of ongoing investigations of our own. We focus upon verb-argument constructions (VACs) in English, such as 'V *across* n' as in "she walked across the street." These initial studies convince us that learners do not acquire language from unstructured, unhelpful experience. Instead, the evidence of language usage is rich in latent structure. Learners' explorations of this problem-space are grounded and contextualized. There is much latent structure to scaffold development in the frequency distributions of exemplars of linguistic constructions and in the network structure of the corresponding semantic space. Furthermore,

learners' investigations of the problem space of language usage are often directed, attentionally focused, and co-constructed in discourse interaction by an interlocutor as a helpful guide, although consideration of these aspects is beyond the scope of this chapter.

Our shared language understanding suggests that, just as for nouns, there is a basic variety of VACs, each with their own basic level verb prototype. For example, despite the fact that you and I have not heard the same input, our experience allows us similar interpretations of novel utterances like "it mandools across the ground" or "the teacher spugged the boy the book." You know that *mandool* is a verb of motion and have some idea of how mandooling works and its action semantics. You know that *spugging* involves transfer, that the teacher is the donor, the boy the recipient, and the book the transferred object. How is this possible, given that you have never previously heard these verbs? Each word of the construction contributes individual meaning, and the verb meanings in these VACs are usually at the core. But the larger configuration of words as a whole also carries meaning. The VAC, as a category, has inherited its schematic meaning from the conspiracy of all of the examples you have heard. *Mandool* inherits its interpretation from the echoes of the verbs that occupy this VAC—words like *come*, *walk*, *move*, . . . , *scud*, *skitter*, and *flit*.

As you read these utterances, you parse them and identify their syntagmatic form: "it mandools across the ground" as a verb locative (VL) construction; "the teacher spugged the boy the book" as a double-object (VOO) construction. Then the paradigmatic associations of the types of verb that fill these slots are awakened: for the VL 'V *across* n' pattern, you associate *mandool* with *come, walk, move, . . . , scud, skitter*, and *flit*; for the VOO construction, you associate *spugged* with *give, send, pass*, . . . , *read, loan,* and *fax*. Knowledge of language is based on these types of inference of syntactic and semantic bootstrapping. Verbs are the cornerstone of the syntax-semantics interface, which is why we focus upon VACs in our research.

In the rest of this chapter we consider the nature of VACs and the psychology of their learning before turning to a complex systems analysis of the dynamic structure of language usage and how it might support robust language learning.

Constructions and Their Acquisition

Construction Grammar

We take the Saussurian (1916) view that the units of language are constructions—form-meaning mappings, conventionalized in the speech community, and entrenched as language knowledge in the learner's mind. They are the symbolic units of language that relate the defining properties of their morphological, lexical, and syntactic form with particular semantic, pragmatic, and discourse functions (Goldberg 1995, 2006). Construction grammar argues that all grammatical phenomena can be understood as learned pairings of form (from morphemes, words, and idioms to partially lexically filled and fully general phrasal patterns) and their associated semantic or discourse functions: "the network of constructions captures our grammatical knowledge *in toto*, i.e., it's constructions all the way down" (Goldberg 2006, 18). Such beliefs, increasingly influential in the study of child language acquisition, emphasize

data-driven, emergent accounts of linguistic systematicities (e.g., Ambridge and Lieven 2011; Ellis 2011; Tomasello 2003).

The Psychology of Learning

Usage-based approaches hold that we learn linguistic constructions while engaging in communication (Bybee 2010). Psycholinguistic research provides the evidence of usage-based acquisition in its demonstrations that language processing is exquisitely sensitive to usage frequency at all levels of language representation—from phonology, through lexis and syntax, to sentence processing (Ellis 2002). Since language users are sensitive to the input frequencies of these patterns, they must have registered their occurrence in processing. Therefore, these frequency effects are compelling evidence for usage-based models of language acquisition that emphasize the role of input. Language knowledge involves statistical knowledge, so humans more easily learn and process high frequency forms and regular patterns, which are exemplified by many types and have few competitors (e.g., Ellis 2006a; MacWhinney 2001).

Constructionist accounts of language learning involve the distributional analysis of the language stream and the parallel analysis of contingent perceptuo-motor activity. Abstract constructions are often learned as categories from the integrated experience of concrete exemplars of usage that follow statistical learning mechanisms relating input and learner cognition (Bybee and Hopper 2001; Christiansen and Chater 2001; Ellis 2006b; Jurafsky and Martin 2009).

Determinants of Construction Learning

Psychological analyses of the learning of constructions as form-meaning pairs are informed by the literature on the associative learning of cue-outcome contingencies. The usual determinants include: (1) form frequency in the input (type-token frequency, Zipfian distribution), (2) function (prototypicality of meaning), and (3) interactions between these (contingency of form-function mapping) (Ellis and Cadierno 2009).

Construction Frequency. Frequency of exposure promotes learning and entrenchment (e.g., Anderson 2000; Ebbinghaus 1885). Learning, memory, and perception are all affected by frequency of usage; the more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization (Harnad 1987; Lakoff 1987; Taylor 1998).

Type and Token Frequency. Token frequency counts how often a particular form appears in the input. The greater the token frequency of an exemplar, the more it contributes to defining the category, and the greater the likelihood it will be considered the prototype. On the other hand, type frequency refers to the number of distinct lexical items that can be substituted in a given slot in a construction, whether it is a word-level construction for inflection or a syntactic construction specifying the relation among words. For example, the regular English past tense *-ed* has a very high type frequency

because it applies to thousands of different types of verbs, whereas the vowel change exemplified in *swam* and *rang* has much lower type frequency. The productivity of phonological, morphological, and syntactic constructions is a function of type rather than token frequency (Bybee and Hopper 2001).

Zipfian Distribution. In natural language, Zipf's law (Zipf 1935) describes how the highest frequency words account for the most linguistic tokens. Zipf's law states that the frequency of words decreases as a power function of their rank in the frequency table. Thus, in English, the most frequent word (*the* with a token frequency of about 60,000 occurrences per million words) occurs approximately twice as often as the second most frequent word, three times as often as the third most frequent word, and so on. If P_f is the proportion of words whose frequency in a given language sample is f, then $P_f \sim f^{-\gamma}$, with $\gamma \approx 1$. Zipf showed that this scaling law holds across a wide variety of language samples. Subsequent research provides support for this law as linguistically universal. Many language events across scales of analysis follow his power law, including words (Evert 2005), collocations (Bannard and Lieven 2009; Solé et al. 2005), formulaic phrases (O'Donnell and Ellis 2009), morphosyntactic productivity (Baayen 2008), grammatical constructs (Ninio 2006; O'Donnell and Ellis 2010), and grammatical dependencies (Ferrer i Cancho and Solé 2001, 2003; Ferrer i Cancho, Solé, and Köhler 2004). Zipfian covering, which involves splitting concepts hierarchically (e.g., animal, canine, dog, retriever, Labrador, and so on) in order to communicate clearly, determines basic categorization, the structure of semantic classes, and the language form-semantic structure interface (Manin 2008; Steyvers and Tennenbaum 2005). Scale-free laws pervade language structure and usage.

However, this does not apply solely to language structure and usage; power law behavior like this has since been shown to apply to a wide variety of structures, networks, and dynamic processes in physical, biological, technological, social, cognitive, and psychological systems of various kinds (e.g., magnitudes of earthquakes, populations of cities, citations of scientific papers, number of hits received by websites, sizes of airline hubs, perceptual psychophysics, memory, categorization, etc.) (Kello et al. 2010; Newman 2005). It has become a hallmark of complex systems theory. It is tempting to think of Zipfian scale-free laws as universals. Complexity theorists suspect them to be fundamental and are beginning to investigate how they might underlie language processing, learnability, acquisition, usage, and change (Beckner et al. 2009; Ellis and Larsen-Freeman 2009b; Ferrer i Cancho and Solé 2001, 2003; Ferrer i Cancho, Solé, and Köhler 2004; Solé et al. 2005).

Various usage-based linguists (e.g., Boyd and Goldberg 2009; Bybee 2008, 2010; Ellis 2008a; Goldberg 2006; Goldberg, Casenhiser, and Sethuraman 2004; Lieven and Tomasello 2008; Ninio 1999, 2006) suspect that it is the Zipfian coming together of linguistic form and function that makes language robustly learnable despite learners' idiosyncratic experience. For example, in first language acquisition, Goldberg, Casenhiser, and Sethuraman (2004) demonstrated that there is a strong tendency for VL, verb object locative (VOL), and double object ditransitive (VOO) VACs to be occupied by one single verb with very high frequency in comparison to

other verbs used, a profile that closely mirrors that of the mothers' speech to their children. They argue that this promotes language acquisition because the low variance sample allows learners to discern what will account for most of the category members, with the category defined later with experience of the full breadth of exemplar types.

Function (Prototypicality of Meaning). Categories have graded structure; some members are better exemplars than others. In the prototype theory of concepts (Rosch and Mervis 1975; Rosch et al. 1976), the prototype as the central ideal is the best example of the category, appropriately summarizing the most representative attributes of a category. As the typical instance of a category, it serves as the benchmark against which less representative instances are classified.

In child language acquisition, a small group of semantically general verbs (e.g., *go, do, make, come*) are learned early (Clark 1978; Goldberg 2006; Ninio 1999; Pinker 1989). Ellis and Ferreira-Junior (2009a) show the same is true of the second language acquisition of VL, VOL, and VOO constructions. These first verbs are prototypical and generic in function (*go* for VL, *put* for VOL, and *give* for VOO). In the early stages of learning categories from exemplars, acquisition might thus be optimized by the introduction of an initial low-variance sample centered upon prototypical exemplars.

Contingency of Form-Function Mapping. Psychological research into associative learning has long recognized that, while frequency of form is important, more so is contingency of mapping (Shanks 1995). For example, when looking at exemplars of birds, though eyes and wings are equally frequently experienced features, wings are the distinctive feature that differentiates birds from other animals. Wings are important features because they are reliably associated with class membership; eyes are a more universal and less reliable trait. Raw frequency of occurrence is less important than the contingency between cue and interpretation (Rescorla 1968). Contingency, reliability of form-function mapping, and associated aspects of predictive value, information gain, and statistical association are driving forces of learning. They are also central in psycholinguistic theories of language acquisition (Ellis 2006a, 2006b, 2008b; Gries and Wulff 2005; MacWhinney 1987).

Usage-Based Acquisition

The primary motivation of construction grammar involves bringing together linguistic form, learner cognition, and usage. Constructions cannot be defined purely on the basis of linguistic form, semantics, or frequency of usage alone. All three factors are necessary in their operationalization and measurement. Psychology theory relating to the statistical learning of categories suggests that constructions are robustly learnable when they are: (1) Zipfian in their type-token distributions in usage, (2) selective in their verb form occupancy, (3) coherent in their semantics, and (4) similar in form and function.

Taking this into account, it is important to measure whether language usage provides experience of this type. If it does, then VACs as linguistic constructions should

be robustly learnable. Is language, which is shaped by the human brain (Christiansen and Chater 2008), consequently shaped *for* the human brain, in that the structures latent in language usage make language robustly learnable?

Language Use as a Complex Adaptive System

This fundamental claim that Zipfian distributional properties of language usage help to make language learnable has just begun to be explored for a small number of VACs in first (Goldberg 2006; Goldberg, Casenhiser, and Sethuraman 2004; Ninio 2006, 2011) and second language acquisition (Ellis and Ferreira-Junior 2009a, 2009b; Ellis and Larsen-Freeman 2009a). It remains a priority to explore its generality across the wide range of English verbal grammar. In order to do this, we need: (1) an integrative analysis of the VACs of a language, the lexis they select, their meanings, and the distributions and mappings of these forms and functions; (2) to determine if the latent structures therein are of the type that would promote robust learning; (3) and to measure corpora of learner language (L1 and L2) to see if learning is shaped by the input. These steps are no small task. Here, we describe some of our pilot work and suggestions for further research.

A Usage-Based Grammar of English Verbs and Verb-Argument Constructions

A Catalogue of VACs. In order to avoid circularity, the determination of the semantic associations of particular linguistic forms should start from structuralist definitions of VACs defined by bottom-up means that are semantics-free. No one in corpus linguistics trusts the text more than Sinclair (2004) in his operationalizations of linguistic constructions on the basis of repeated patterns of words in collocation, colligation, and phrases. We therefore use the patterns presented in the *Grammar Patterns: Verbs* (Hunston and Francis 1996) that arose out of the COBUILD project (Sinclair 1987) for our initial analyses. There are over seven hundred patterns of varying complexity in this volume. The form-based patterns described in the COBUILD Verb Patterns volume (Francis, Hunston, and Manning 1996) take the form of word class and lexis combinations, such as the 'V *across* n' pattern:

> The verb is followed by a prepositional phrase which consists of *across* and a noun group.
>
> This pattern has one structure:
>
> * Verb with Adjunct.
>
> *I cut across the field.*

Our initial research (for further details see Ellis and O'Donnell 2012; Römer, O'Donnell, and Ellis, in press) describes the methods and findings for an initial convenience sample of twenty-three VACs, most of which follow the verb-preposition-noun phrase structure, such as 'V *into* n', 'V *after* n', 'V *as* n' (Goldberg 2006). We also include other classic examples, such as the 'V n n' ditransitive and the *way* construction.

A Large Corpus of English. To get a representative sample of usage, one needs a large corpus. We investigate the verb type-token distribution of these VACs in the one hundred million word British National Corpus (BNC 2007), parsed using the XML version of the BNC and the RASP parser (Briscoe, Carroll, and Watson 2006). For each VAC, we translate the formal specifications from the COBUILD patterns into queries to retrieve instances of the pattern from the parsed corpus. Using a combination of part of speech, lemma, and dependency constraints, we formulate queries for each of the construction patterns. For example, the 'V *across* n' pattern is identified by looking for sentences that have a verb form within three words of an instance of *across* as a preposition, where there is an indirect object relation holding between *across* and the verb, and where the verb does not have any other object or complement relations to following words in the sentence.

The Lexical Constituency of Verbs in VACs. The sentences extracted using this procedure produced verb type frequency distributions like the following one for the 'V *across* n' VAC:

come	483				
walk	203				
cut	199	...			
run	175	veer	4		
...		slice	4	...	
		...		navigate	1
				scythe	1
				scroll	1

These distributions appear to be Zipfian, exhibiting the characteristic long-tail in a plot of rank against frequency. We generated logarithmic plots and linear regressions to examine the extent of this trend using logarithmic binning of frequency against log cumulative frequency following Adamic and Huberman (2002). Figure 3.1 shows such a plot for verb type frequency of the 'V *across* n' construction alongside the same plot for verb type frequency of the ditransitive 'V n n' construction. In these graphs, we randomly selected one verb from each frequency bin for illustration. Both distributions produce a good fit of Zipfian type-token frequency with $r^2 > 0.97$ and slope (γ) around 1. Inspection of the construction verb types, ranked from most frequent to least frequent, also demonstrates that the lead member is prototypical of the construction and generic in its action semantics.

Since Zipf's law applies across language, the Zipfian nature of these distributions is potentially trivial. But they are more interesting if the company of verb forms occupying a construction is selective—in other words, if the frequencies of the particular VAC verb members cannot be predicted from their frequencies in language as a whole. We measure the degree to which VACs are selective by using measures of association and the statistic '1-tau', where Kendall's tau measures the correlation between the rank verb frequencies in the construction and in language as a whole. For the VACs studied so far, 1-tau is typically about 0.7, showing that the rankings of

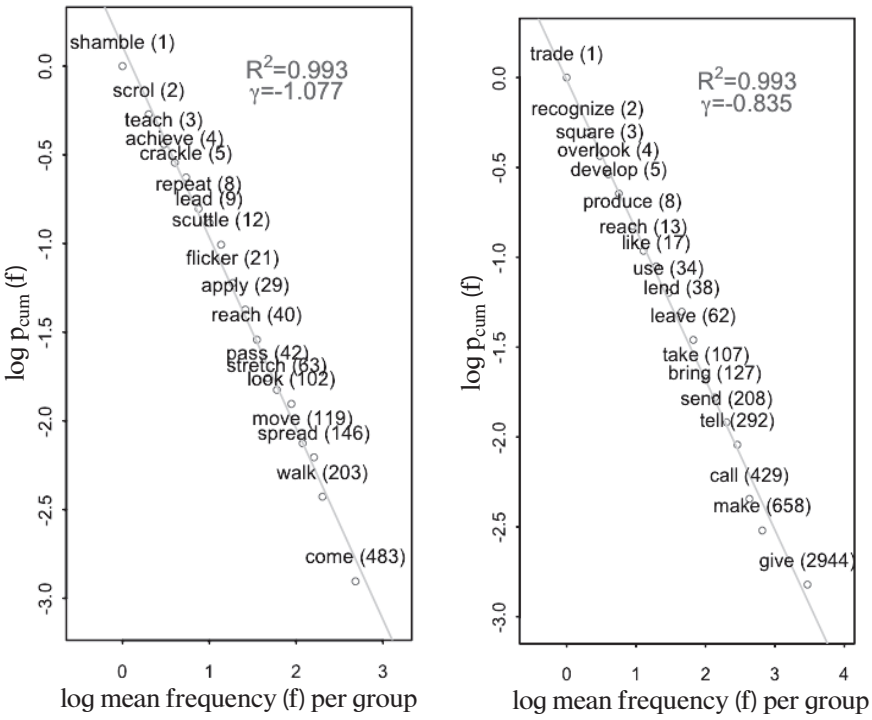

Figure 3.1 BNC Verb Type Distribution for 'V *across* n' and for 'V n n'

verbs in particular VACs differ markedly from the rankings of verbs in the language as a whole. VACs are selective in their verb constituency.

Verb-VAC Contingency. Another way of looking at this is to assess verb-VAC contingency. Some verbs are closely tied to a particular construction; for example, *give* is highly indicative of the ditransitive construction, whereas *leave*, although it can form a ditransitive, is more often associated with other constructions, such as the simple transitive or intransitive. The more reliable the contingency between a cue and an outcome, the more readily an association between them can be learned (Shanks 1995). Thus, constructions with more faithful verb members should be more readily acquired and higher contingency verbs should be learned first in a VAC and should come to mind first when processing that VAC. The measures of contingency we adopt are: (1) faithfulness—the proportion of tokens of total verb usage that appear in this particular construction (e.g., the faithfulness of *give* to the ditransitive is approximately 0.40; that of *leave* is 0.01); (2) directional mutual information (MI Word → Construction: *give* 16.26, *leave* 11.73 and MI Construction → Word: *give* 12.61 *leave* 9.11); and (3) the one-way dependency statistic ΔP (Allan 1980) used in the associative learning literature (Shanks 1995) and in other studies of form-function contingency in construction usage, knowledge, and processing (Ellis 2006a;

Ellis and Ferreira-Junior 2009b; Ellis, O'Donnell, and Römer, in press). Our analyses for the twenty-three VACs studied so far show a general pattern in which individual verbs tend to select particular constructions (MI_{wc}, ΔP_{wc}) and particular constructions select particular verbs (MI_{cw}, ΔP_{cw}) (for details see Ellis and O'Donnell 2012; Ellis, O'Donnell, and Römer, in press).

VAC Meanings and Coherence. Our semantic analyses use WordNet, a distribution-free semantic database based on psycholinguistic theory that has been in development since 1985 (Miller 2009). WordNet places words into a hierarchical network. At the top level, the hierarchy of verbs is organized into 559 distinct root synonym sets (synsets such as *move1* expressing translational movement, *move2* expressing movement without displacement, etc.) which is then split into over 13,700 verb synsets. Verbs are linked in the hierarchy according to relations such as hypernym (verb Y is a hypernym of the verb X if the activity X is a type of Y—to *perceive* is a hypernym of to *listen*), and hyponym (verb Y is a hyponym of the verb X if the activity Y is doing X in some manner—to *lisp* is a hyponym of to *talk*). Various algorithms to determine the semantic similarity between WordNet synsets have been developed that consider the distance between the conceptual categories of words, as well as the hierarchical structure of the WordNet (Pedersen, Patwardhan, and Michelizzi 2004). We take the lists of verbs occupying each VAC using the methods described earlier and compare the verbs pairwise on these metrics. We then apply networks science, such as graph-based algorithms (de Nooy, Mrvar, and Batagelj 2010) to build semantic networks in which the nodes represent verb types and the edges show strong semantic similarity for each VAC. Standard measures of network density, average clustering, degree centrality, transitivity, and so on are used to assess the cohesion of these semantic networks; we also apply algorithms for the detection of communities within the networks representing different semantic sets (Clauset, Newman, and Moore 2004; Danon et al. 2005). The network for 'V *across* n' is shown as an example in figure 3.2. The network is fairly dense. The hubs, shown here as larger nodes, are those that are most connected, meaning they have the highest degree. They are *go*, *move*, and *travel*—the prototypical 'V *across* n' senses. However, there are also subcommunities, shown in different colors. For example, one relating to vision includes *look*, *stare*, *gaze*, *face*; another relating to speeded movement with unspecified action semantics includes *shoot*, *skud*, *race*, *rush*, etc.; and another emphasizing flat contact includes *lay*, *lie*, *sprawl*, etc. Note that degree in the network is unrelated to token frequency in the corpus; it simply reflects verb type connectivity within the network. This also applies to betweenness centrality, which was developed as a measure quantifying the control of a human on the communication between other humans in a social network. Betweenness centrality is a measure of a node's centrality in a network equal to the number of shortest paths from all vertices to all others that pass through that node. In semantic networks, central nodes are those which are prototypical of the network as a whole.

Across the VACs we have investigated to date (O'Donnell, Ellis, and Corden 2012), the semantic networks are coherent, with short path-lengths between the nodes and degree distributions which approximate a Zipfian power function. Satisfaction of

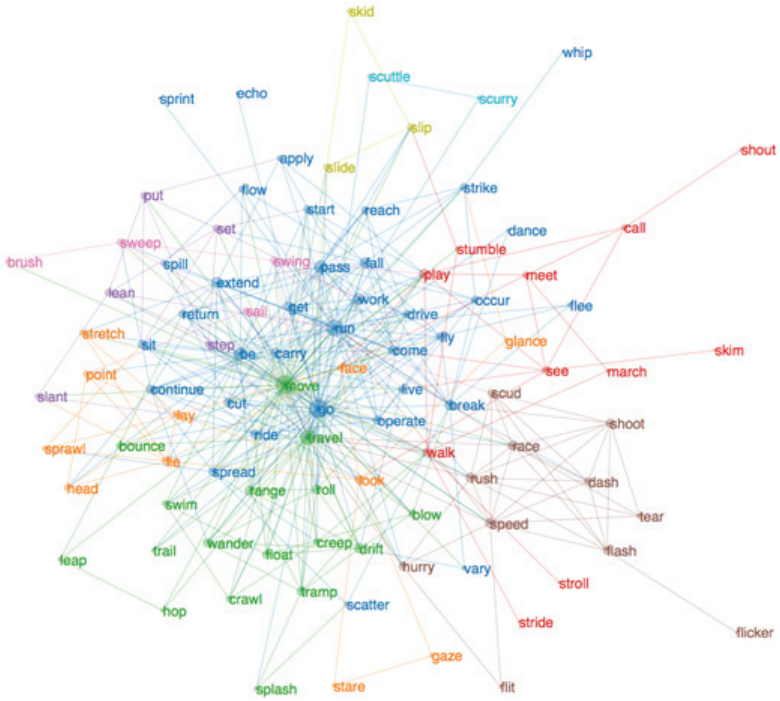

Figure 3.2 A Semantic Network for 'V *across* n' from the BNC Using WordNet as a Base

these two criteria classifies them as small-world networks (Barabási 2002). Steyvers and Tennenbaum (2005) show more generally across semantic networks—such as Roget's thesaurus and WordNet—that the distributions of the number of connections as a whole follow power laws that indicate a scale-free pattern of connectivity, with most nodes having relatively few connections joined together through a small number of hubs with many connections. Small-world networks are interesting because they are more robust to perturbations than other network architectures (Barabási 2002). They achieve this through the high connectivity of their hubs.

Is the Latent Structure of VACs of the Type That Would Promote Robust Acquisition following the Proposed Learning Principles?

Our core research questions concern the degree to which VAC form, function, and usage might promote robust learning. As we explained, the psychology of learning as it relates to these psycholinguistic matters suggests, in essence, that learnability will be optimized for constructions that are: (1) Zipfian in their type-token distributions in usage, (2) selective in their verb form occupancy, and (3) coherent in their semantics. The findings we summarized earlier confirm that these factors apply.

Remember that the two Zipfian distributions described earlier are different. The first relates to type-token frequency distribution in the language. The second relates to node connectivity (degree distribution and betweenness centrality) in the semantic network, and it has no regard of corpus frequencies. Nevertheless, the high-degree items in the semantic distribution also tend to be the high-token frequency items in the corpus. We believe that it is this coming-together of the two Zipfian distributions that makes language robustly learnable. The VAC pattern is seeded by a high-token frequency exemplar that is also prototypical in the action-dynamic construal of events to which it relates. Thereafter, the forms and functions of verbs added to the VACs resonate with the category itself. "(T)his process is actually the organization of exemplars of utterances and of verb-specific constructions into clusters of greater or lesser size, with greater or lesser semantic coherence" (Croft 2012, 393).

Further Directions and Conclusions

These initial investigations make it clear that usage is intricately structured in ways typical of complex adaptive systems in that there are scale-free distributions in verb usage frequency within constructions and scale-free connectivity patterns within semantic networks. This latent structure could potentially scaffold robust development.

Note that these results are preliminary, as they are based on an analysis of only about twenty constructions to date. There are over seven hundred patterns of varying complexity in the *Grammar Patterns #1: Verbs* (Francis, Hunston, and Manning 1996).

There is a considerable amount of additional statistical analysis and modeling to be done, as well as analyses of longitudinal corpora of language acquisition (and the matching child directed speech or NS interlocutor language) to test out the predictions of learning. Experimental psycholinguistic research is also necessary to test the psychological validity of VACs.

We identify the following priorities:

- Learning theory makes *relative* predictions as well, and these should inform our understanding of language acquisition and processing. VACs that are higher frequency in the language, with more verb types and greater semantic cohesion, should be acquired earlier and accessed more readily in speakers' minds. Similarly, verbs that are more frequent constituents of a VAC, more faithful to that VAC, and closer to the VAC semantic prototype should be acquired earlier in that VAC and accessed more readily in speakers' minds when they consider that VAC schema. A considerable amount of statistics and modeling remains necessary to test these hypotheses. We report some initial findings in Ellis, O'Donnell, and Römer (in press).
- This work needs to be done for second language acquisition, too. We have made a start (Ellis and Ferreira-Junior 2009a, 2009b), but there is an imperative for larger L2 corpora.
- The psychological reality of VACs needs further work. In various studies we use free association tasks to have people think of the first word that comes to

mind to fill the V slot in a particular VAC frame. The responses of adult native and fluent L2 learners are highly predicted by corpus frequencies, showing that users have implicit knowledge of the occupancy and frequencies of verbs in VACs (Ellis and O'Donnell 2011; Römer, O'Donnell, and Ellis, in press). We are particularly excited that we have identified separable influences—(1) the frequency of the verb in the language, (2) VAC-verb contingency (ΔP_{cw}), and (3) the prototypicality of the verb in the semantic network (as indexed by its betweenness centrality)—as factors in the determination of the verbs freely associated with skeletal VAC frames.

- The semantic analyses here are crude; other distributional measures could be well applied alongside techniques for investigating network growth (O'Donnell, Ellis, and Corden 2012).
- The acquisition data here are basically correlational. There need to be experimental studies comparing the relative learnability of Zipfian skewed input compared with languages with flatter frequency profiles. Casenhiser and Goldberg (2005) and Goldberg, Casenhiser, and Sethuraman (2004) have made important steps in doing this in children and adults, but they investigate the learning of just one construction from a small number of trials, and there is need for larger causal studies of the effects of combined Zipfian frequency distributions and Zipfian semantic connectivity upon more complete approximations to natural language.
- To better understand the processes of how these latent structures of usage affect robust acquisition and stable usage, there is need for modeling, both connectionist (Ellis and Larsen-Freeman 2009a) and agent-based (Beckner et al. 2009).

Meanwhile, we can at least say that the input from which learners acquire language is far from unstructured, unhelpful, or barren. The evidence of language usage is rich in latent structure. We believe that, with language as with other cognitive realms, our experiences conspire to give us competence. In the spirit of GURT 2012, such beliefs become either firmer or falsified, dependent upon the extent to which we can measure and test them.

ACKNOWLEDGMENTS

We thank Katie Erbach, Mary Smith, Lucy Zhao, Gin Corden, Danny Tsu-yu Wu, Liam Considine, Jerry Orlowski, and Sarah Garvey for help in the design, data collection, and analysis of these data. We also thank the University of Michigan LSA Scholarship/Research funding opportunity for supporting the project "Piloting the development of an inventory of usage of English verb grammar."

REFERENCES

Adamic, L. A., and B. A. Huberman. 2002. "Zipf's law and the Internet." *Glottometrics* 3: 143–50.

Allan, L. G. 1980. "A note on measurement of contingency between two binary variables in judgment tasks." *Bulletin of the Psychonomic Society* 15: 147–49.

Ambridge, B., and E. Lieven. 2011. *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge: Cambridge University Press.

Anderson, J. R. 2000. *Cognitive Psychology and Its Implications*, fifth edition. New York: W. H. Freeman.

Baayen, R. H. 2008. "Corpus linguistics in morphology: morphological productivity." In *Corpus Linguistics: An International Handbook*, edited by A. Ludeling and M. Kyto. Berlin: Mouton De Gruyter.

Bannard, C., and E. Lieven. 2009. "Repetition and reuse in child language learning." In *Formulaic Language Volume Two: Acquisition, Loss, Psychological Reality, and Functional Explanations*, 299–321, edited by R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley. Amsterdam: John Benjamins.

Barabási, A. L. 2002. *Linked: The New Science of Networks.* Cambridge, MA: Perseus Books.

Beckner, C., R. Blythe, J. Bybee, M. H. Christiansen, W. Croft, N. C. Ellis, J. Holland, J. Ke, D. Larsen-Freeman, and T. Schoenemann. 2009. "Language is a complex adaptive system." Position paper. *Language Learning, 59 Supplement 1*, 1–26.

BNC. 2007. BNC XML Edition. www.natcorp.ox.ac.uk/corpus/.

Boyd, J. K., and A. E. Goldberg. 2009. "Input effects within a constructionist framework." *Modern Language Journal* 93 (2): 418–29.

Briscoe, E., J. Carroll, and R. Watson. 2006. "The second release of the RASP system." Paper presented at the Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia.

Bybee, J. 2008. "Usage-based grammar and second language acquisition." In *Handbook of Cognitive Linguistics and Second Language Acquisition*, edited by P. Robinson and N. C. Ellis. London: Routledge.

———. 2010. *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.

Bybee, J., and P. Hopper, editors. 2001. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.

Casenhiser, D., and A. E. Goldberg. 2005. "Fast mapping between a phrasal form and meaning." *Developmental Science* 8: 500–508.

Christiansen, M. H., and N. Chater. 2008. "Language as shaped by the brain." *Behavioral & Brain Sciences* [target article for multiple peer commentary] 31: 489–509.

Christiansen, M. H., and N. Chater, editors. 2001. *Connectionist Psycholinguistics*. Westport, CO: Ablex.

Clark, E. V. 1978. "Discovering what words can do." In *Papers from the Parasession on the Lexicon,* 34–57, edited by D. Farkas, W. M. Jacobsen, and K. W. Todrys. Chicago Linguistics Society, April 14-15, 1978. Chicago: Chicago Linguistics Society.

Clauset, A., M. E. J. Newman, and C. Moore. 2004. "Finding community structure in very large networks." *Physical Review E* 70: 066111.

Croft, W. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective.* Oxford: Oxford University Press.

———. 2012. *Verbs: Aspect and Causal Structure*. Oxford: Oxford University Press.

Danon, L., A. Díaz-Guilera, J. Duch, and A. Arenas. 2005. "Comparing community structure identification methods." *Journal of Statistical Mechanics* 29: P09008.

de Nooy, W., A. Mrvar, and V. Batagelj. 2010. *Exploratory Social Network Analysis with Pajek*. Cambridge: Cambridge University Press.

Ebbinghaus, H. 1885. *Memory: A Contribution to Experimental Psychology*. Translated by H. A. R. C. E. B.: 1913. New York: Teachers College, Columbia.

Ellis, N. C. 1998. "Emergentism, connectionism and language learning." *Language Learning* 48 (4): 631–64.

———. 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): 143–88.

———. 2006a. "Language acquisition as rational contingency learning." *Applied Linguistics,* 27 (1): 1–24.

———. 2006b. "Selective attention and transfer phenomena in SLA: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning." *Applied Linguistics* 27 (2): 1–31.

———. 2008a. "Optimizing the input: Frequency and sampling in usage-based and form-focused learning." In *Handbook of Second and Foreign Language Teaching*, edited by M. H. Long and C. Doughty. Oxford: Blackwell.

———. 2008b. "Usage-based and form-focused language acquisition: The associative learning of constructions, learned-attention, and the limited L2 endstate." In *Handbook of Cognitive Linguistics and Second Language Acquisition*, 372–405, edited by P. Robinson and N. C. Ellis. London: Routledge.

———. 2011. "The emergence of language as a complex adaptive system." In *Handbook of Applied Linguistics*, 666–79, edited by J. Simpson. London: Routledge/Taylor Francis.

Ellis, N. C., and T. Cadierno. 2009. "Constructing a second language." *Annual Review of Cognitive Linguistics* 7 (special section): 111–290.

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

———. 2009b. "Constructions and their acquisition: Islands and the distinctiveness of their occupancy." *Annual Review of Cognitive Linguistics*: 111–39.

Ellis, N. C., and D. Larsen-Freeman. 2006. "Language emergence: implications for applied linguistics." *Applied Linguistics* 27 (4).

———. 2009a. "Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage." *Language Learning* 59 (Supplement 1): 93–128.

———. 2009b. "Language as a complex adaptive system (special issue)." *Language Learning* 59 (Supplement 1).

Ellis, N. C., and M. B. O'Donnell. 2011. "Robust language acquisition—an emergent consequence of language as a complex adaptive system." In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 3512–17, edited by L. Carlson, C. Hölscher, and T. Shipley. Austin, TX: Cognitive Science Society.

———. 2012. "Statistical construction learning: Does a Zipfian problem space ensure robust language learning?" In *Statistical Learning and Language Acquisition*, 265–304, edited by J. Rebuschat and J. Williams. Berlin: Mouton de Gruyter.

Ellis, N. C., M. B. O'Donnell, and U. Römer. 2012. "Usage-Based language: Investigating the latent structures that underpin acquisition." *Currents in Language Learning* 1, 63: Suppl. 1, pp. 25–51.

———. In press. "The processing of verb-argument constructions is sensitive to form, function, frequency, contingency, and prototypicality." *Cognitive Linguistics* 25 (1).

Evert, S. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Stuttgart: University of Stuttgart.

Ferrer i Cancho, R., and R. V. Solé. 2001. "The small world of human language." *Proceedings of the Royal Society of London* 268: 2261–65.

———. 2003. "Least effort and the origins of scaling in human language." *PNAS* 100: 788–91.

Ferrer i Cancho, R., R. V. Solé, and R. Köhler. 2004. "Patterns in syntactic dependency networks." *Physical Review* E69: 0519151–58.

Francis, G., S. Hunston, and E. Manning, editors. 1996. *Grammar Patterns 1: Verbs. The COBUILD Series*. London: Harper Collins.

Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure.* Chicago: University of Chicago Press.

———. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldberg, A. E., D. M. Casenhiser, and N. Sethuraman. 2004. "Learning argument structure generalizations." *Cognitive Linguistics* 15: 289–316.

Gries, S. T., and S. Wulff. 2005. "Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora." *Annual Review of Cognitive Linguistics* 3: 182–200.

Harnad, S, editor. 1987. *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Holland, J. H. 1995. *Hidden Order: How Adaption Builds Complexity*. Reading: Addison-Wesley.

Hunston, S., and G. Francis. 1996. *Pattern Grammar: A Corpus Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Jurafsky, D., and J. H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, second edition. Englewood Cliffs, NJ: Prentice Hall.

Kello, C. T., G. D. A. Brown, R. Ferrer i Cancho, J. G. Holden, K. Linkenkaer-Hansen, T. Rhodes, and G. C. Van Orden. 2010. "Scaling laws in cognitive sciences." *Trends in Cognitive Science* 14: 223–32.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Langacker, R. W. 1987. *Foundations of Cognitive Grammar: Volume One. Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Larsen-Freeman, D. 1997. "Chaos/complexity science and second language acquisition." *Applied Linguistics* 18: 141–65.

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Analysis*. Chicago: Chicago University Press.

Lieven, E., and M. Tomasello. 2008. "Children's first language acquisition from a usage-based perspective." In *Handbook of Cognitive Linguistics and Second Language Acquisition*, edited by P. Robinson and N. C. Ellis. New York and London: Routledge.

MacWhinney, B. 1987. "The competition model." In *Mechanisms of Language Acquisition*, 249–308, edited by B. MacWhinney. Hillsdale, NJ: Erlbaum.

———. 2001. "The competition model: The input, the context, and the brain." In *Cognition and Second Language Instruction*, 69–90, edited by P. Robinson. New York: Cambridge University Press.

MacWhinney, B, editor. 1999. *The Emergence of Language*. Hillsdale, NJ: Erlbaum.

Manin, D. Y. 2008. "Zipf's law and avoidance of excessive synonymy." *Cognitive Science* 32: 1075–98.

Miller, G. A. 2009. "WordNet—about us." Princeton University. Accessed March 1, 2010. http://wordnet.princeton.edu

Newman, M. 2005. "Power laws, Pareto distributions and Zipf's law." *Contemporary Physics* 46: 323–51.

Ninio, A. 1999. "Pathbreaking verbs in syntactic development and the question of prototypical transitivity." *Journal of Child Language* (26): 619–53.

———. 2006. *Language and the Learning Curve: A New Theory of Syntactic Development*. Oxford: Oxford University Press.

———. 2011. *Syntactic Development, Its Input and Output*. Oxford: Oxford University Press.

O'Donnell, M. B., and N. C. Ellis. 2009. "Measuring formulaic language in corpora from the perspective of language as a complex system." Paper presented at the Fifth Corpus Linguistics Conference, University of Liverpool, July 20–23.

———. 2010. "Towards an inventory of English verb argument constructions." Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles.

O'Donnell, M. B., N. C. Ellis, and G. Corden. 2012. "Exploring semantics in verb argument constructions using community identification algorithms." Paper presented at the Language and Network Science Symposium at the International Conference on Network Science NETSCI 2012, Northwestern University.

Page, S. E. 2009. Understanding Complexity [DVD-ROM]. Chantilly, VA: The Teaching Company.

Pedersen, T., S. Patwardhan, and J. Michelizzi. 2004. "WordNet: Similarity—Measuring the relatedness of concepts." Paper presented at the Proceedings of Fifth Annual Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL 2004).

Pinker, S. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: Bradford Books.

Rescorla, R. A. 1968. "Probability of shock in the presence and absence of CS in fear conditioning." *Journal of Comparative and Physiological Psychology* 66: 1–5.

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

Römer, U., M. O'Donnell, and N. C. Ellis. In press. "Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions: Exploring corpus data and speaker knowledge." In *Corpora, Grammar, Text and Discourse: In Honour of Susan Hunston*, edited by M. Charles, N. Groom, and S. John. Amsterdam: John Benjamins.

Rosch, E., and C. B. Mervis. 1975. "Cognitive representations of semantic categories." *Journal of Experimental Psychology: General* 104: 192–233.

Rosch, E., C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. 1976. "Basic objects in natural categories." *Cognitive Psychology* 8: 382–439.

Saussure, F. D. 1916. *Cours de Linguistique Générale*, translated by Roy Harris. London: Duckworth.

Shanks, D. R. 1995. *The Psychology of Associative Learning*. New York: Cambridge University Press.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Sinclair, J, editor. 1987. *Looking Up: An Account of The COBUILD Project in Lexical Computing*. London: Collins ELT.

Solé, R. V., B. Murtra, S. Valverde, and L. Steels. 2005. Language Networks: Their Structure, Function and Evolution. *Trends in Cognitive Sciences* 12.

Steyvers, M., and J. Tennenbaum. 2005. "The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth." *Cognitive Science* 29: 41–78.

Talmy, L. 2000. *Toward a Cognitive Semantics: Concept Structuring Systems*. Cambridge, MA: MIT Press.

Taylor, J. R. 1998. "Syntactic constructions as prototype categories." In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, 177–202, edited by M. Tomasello. Mahwah, NJ: Erlbaum.

———. 2002. *Cognitive Grammar*. Oxford: Oxford University Press.

Tomasello, M. 2003. *Constructing a Language*. Boston, MA: Harvard University Press.

Zipf, G. K. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: The MIT Press.