LiLT volume 11, issue 1 December 2014

What is grammar like? A usage-based constructionist perspective

Vsevolod Kapatsinski

This paper is intended to elucidate some implications of usage-based linguistic theory for statistical and computational models of language acquisition, focusing on morphology and morphophonology. I discuss the need for grammar (a.k.a. abstraction), the contents of individual grammars (a potentially infinite number of constructions, paradigmatic mappings and predictive relationships between phonological units), the computational characteristics of constructions (complex non-crossover interactions among partially redundant features), resolution of competition among constructions (probability matching), and the need for multimodel inference in modeling internal grammars underlying the linguistic performance of a community.

1 Introduction

Usage-based linguistics is a relatively recent approach to linguistic theory[1] that has rapidly risen in prominence in the last two decades. Like most approaches to linguistic theory, usage-based linguistics is interested in explaining why languages are the way they are. Usage-based linguists take a dynamic approach to explanation: what we seek to explain are the patterns of language change, and we take the true universals of language to be the cognitive and social mechanisms responsible for language change; see Bybee (2001):189-215.[2]

[1] The term itself dates back to Langacker (1987).

[2] For example, Bybee (2003) tries to explain the universal diachronic process of grammaticalization, whereby lexical items (like *going to* in the sense of locomo-

LiLT Volume 11, December 2014.
Theoretical and Computational Morphology: New Trends and Synergies.

Unlike classical generative linguistics, usage-based linguistics is empiricist in its approach to language acquisition.[3] We think that it is more productive to follow the working assumption that linguistic knowledge is learned and try to figure out how it could be learned, rather than to assume a rich innate store of linguistic knowledge; Bybee (2001):212, Tomasello (2003), see also Hayes and Wilson (2008)'s call for a *learning-theoretic* phonology; cf. Chomsky (1981, 1986) for the opposite view.

Like in generative linguistics, e.g. Chomsky and Halle (1965), mechanisms of language acquisition and biases inherent to these mechanisms are an important locus of explanation for why languages are the way they are, and how they are likely to change. However, in usage-based linguistics, acquisition biases are not the *only* locus of explanation. For example, a prominent place in the usage-based linguist's arsenal of explanatory mechanisms is reserved for articulatory ease, e.g. Bybee (2001, 2003, 2006); Browman and Goldstein (1992); Hooper (1976); Mowrey and Pagliuca (1995), and perceptual distinctiveness, explored in Liljencrants and Lindblom (1972); Baese-Berk and Goldrick (2009) and Wedel et al. (2013). These biases are assumed to operate throughout one's lifetime in every instance of communication rather than being

tion with the intent to do something) become grammatical items (*gonna*, a future marker). Bybee notes that grammaticalization is accompanied by an increase in frequency of use, as well as phonological and semantic changes. She argues that the changes could be accounted for by the cognitive mechanisms of 1) automatization of production of frequently used units of execution (see Kapatsinski (2010a) for empirical evidence), which causes reduction of the frequently used item, 2) habituation, e.g. Harris (1943), which weakens the connection between the frequently used item and the evoked meaning, and 3) association formation, where frequent contextual inferences become associated with the item that frequently occurs in that context. Increased frequency feeds these changes but is also fed by them, driving the process onward. For example, the grammaticalizing item comes to have a more general meaning (via habituation), which then makes it usable in more contexts, driving further increases in frequency. It also becomes easier to pronounce (via automatization), which makes it more likely to win the competition for production against harder-to-pronounce competitors, increasing its frequency in the future; Martin (2007).

[3] I say *classical* because the current generative position on the issue is rather confusing. Chomsky (1993) makes a radical break with previous generative work in assuming a very minimal "narrow UG", the part of the hypothesized store of innate universal knowledge that is specific to language. Developing this position, Hauser et al. (2002) argue that the only innate knowledge specific to language is the principle of recursion. Despite the radical shift in the theory, generativist grammatical descriptions continue using universal deep structure representations that are then transformed into language-particular surface structures. A universal deep structure fit well with the theory that we are born with a rich store of knowledge about language, as in Chomsky (1981). If recursion is all that is innate and specific to language, the motivation for a universal deep structure is unclear.

specific to children acquiring the basics of their native language. An important role is also ascribed to social dynamics responsible for pattern conventionalization and propagation through the community; Labov (2001); Yu (2010).

Like generative linguistics, usage-based linguistics is mentalist, in that we are interested in the mental representations that allow people to produce and comprehend language, and in the way these mental representations change as a result of experience.[4] However, usage-based linguistics recognizes that, if linguistic theory is to explain why languages are the way they are, we need to be able to account for the interplay between *E-Language* (linguistic behavior) and *I-Language* (the system of mental representations generating this behavior).[5]

In particular, observable behavior is the locus of conventionalization: the target of language acquisition is not a system of mental representations but rather a system of observable behavioral patterns, which are conventionalized at the level of the speech community, as argued by sociolinguists: Labov (1975, 1996), and Weinreich et al. (1968).[6] Mental representations, not being directly observable, are not subject to conventionalization and are therefore free to vary as long as the right behavioral patterns are produced. Behavior patterns that are conventionalized at the community level and thus act as targets in the process of language acquisition, need to be robust enough to be easily transmittable and shared by people with different lexica and different processing styles; Deacon (1997); Mielke et al. (2010); Pierrehumbert (2001). Processes of conventionalization are another important influence on the structures of human languages. Not only would patterns that are not

[4] For example, Bybee (2006):711 writes that "While all linguists are likely to agree that grammar is the cognitive organization of language, a usage-based theorist would make the more specific proposal that grammar is the cognitive organization of one's experience with language." Cf. Householder (1966):100, responding to Chomsky and Halle (1965): "A linguist who could not devise a better grammar than is present in any speaker's brain ought to try another trade".

[5] The terms *I-Language* and *E-Language* are from Chomsky (1986). For the position that I-Language is of particular interest to linguistics within the generative paradigm, see Chomsky and Halle (1965) and Chomsky (1986). For the view that E-Language is central to linguistics, see Bloomfield (1926); Goldsmith (To appear); Householder (1966). While sociolinguistics is often seen as being concerned exclusively with E-Language, e.g. Kay and McDaniel (1979), see Sankoff and Labov (1979) for a more nuanced position.

[6] To the extent that behavior is unobservable, it is not subject to conventionalization. For example, there are two perceptually near-equivalent but articulatorily very different ways to produce the English /ɹ/ (by flexing the tip of the tongue upward or bunching the tongue body). Individual speakers appear to have consistent patterns of /ɹ/ production with no sociolinguistic consequences, which allows the behavioral variation to persist; see Mielke et al. (2010).

robustly transmittable be lost, but also some individuals are in a better position to spread innovations; Labov (2001). Furthermore, the factors that make people likely to spread innovations (such as good social skills and being old enough, young enough, and 'cool' enough to be emulated) are also correlated with processing differences that have implications for the types of innovations they are likely to make. See Bybee (2001):201-203 for innovations that are often made by children but do not appear to spread through the community, and Yu (2010) for implications of a correlation between phonological processing differences and position on the autism spectrum for sound change.

Usage-based constructionist approaches assume that grammar acquisition involves statistical inference, that grammar is stochastic in nature, and ultimately learnable with little a priori knowledge. These assumptions make them highly compatible with statistical models that dominate computational linguistics. The results of usage-based work on a wide variety of languages also appear to have fundamental implications for the plausibility of various model types. However, these implications may not be apparent to those interested in statistical modeling, as work on grammatical theory and statistical inference often uses different terminology. The present paper is intended to make the relations between usage-based linguistic theory and statistical modeling explicit and to highlight both areas where there seems to be consensus within usage-based linguistic theory and areas where more work is needed. Of course, the impression of consensus is just that, an impression, based largely on not having encountered disagreement in the literature or in discussing these issues with other community members. I do not intend to try to speak for all linguists who consider themselves usage-based, nor have I conducted a scientific poll on the issues discussed below. This is no more than an individual variant of the usage-based position. I may be very wrong about the existence of community consensus on some issue. Keep a salt shaker handy.

2 Storage vs. computation and the need for inference (a.k.a. grammar)

What is grammar? In the most general terms, we can say that a grammar is a system of generalizations that subserves linguistic creativity.[7]

[7] *Linguistic creativity* refers to the fact that speakers of a language can produce utterances that they have never experienced that are nonetheless acceptable to other speakers from the same speech community; Chomsky (1975):61. No human language learner assumes that only the utterances s/he experienced are acceptable, and that no other utterances can be produced. In morphology, creativity (also called *productivity*) manifests itself as the ability to produce new forms (or derivations) of

Under a usage-based view of grammar, the grammar is induced from language experience. However, the need for induction and generalization is controversial. There is a sizeable group of researchers who believe in a lazy-learning view of language acquisition, also sometimes called *analogical* or *exemplar-based*, e.g. Arndt-Lappe (2011), Eddington (2000); Goldinger (1998); Skousen (1989). [8] On this view, all there is to language acquisition is memorization of experienced utterances, and no generalization *during acquisition* is in principle necessary. Generalization is done only on an as-needed basis. By contrast, grammatical theories propose that language learning is not lazy: language learners keep track of co-occurrences among features of linguistic stimuli, learning an intricate web of predictive dependencies (perhaps, so that they can cope with a noisy environment).[9] Since this paper is about characteristics of grammars, I will spend some time justifying why we need grammars or, in other words, why the lazy-learning view of language acquisition is inadequate, and why lazy-learning models are nonetheless often successful.

Usage-based linguists differ from generativists in assuming that, as well stated by Householder (1966):100, "table look-up rather than al-

words to express an intended meaning. For example, as famously shown by Berko (1958), knowing that a certain creature is called a *wug*, an English speaker could produce the never-before-encountered plural form *wugs* (and a Russian would produce *wugi* or maybe *wuga*). Given a novel adjective *blig*, an English speaker could say that the degree of being *blig* would be called *bligness*, and a Russian borrowing *blig* from English would convert it into *bligij*, *blignyj*, *bligovyj*, *bligskij*, or *bliguchij* to fit one of the Russian adjectival constructions.

[8] The terminology is somewhat confusing in that the most successful and widely-used analogical models, the Tilburg Memory-Based Learner - *TiMBL*, described in Daelemans and van den Bosch (2005), and the Generalized Context Model as described in Nosofsky (1986), in fact weigh features by their predictive power, and thus are not pure lazy learning models, cf. Hintzman (1986); Skousen (1989) and the 'crippled' version of the Generalized Context Model used in Albright and Hayes (2003). Daelemans et al. (1999) further show that the generalizations acquired by the learner can be expressed as a conditional inference tree incorporating feature weighting. As noted by Baayen et al. (2013b), this allows TiMBL, in contrast with Skousen (1989)'s Analogical Modeling of Language, to avoid the combinatorial explosion that comes from explicitly encoding all exemplars separately and therefore to handle realistically detailed linguistic representations.

[9] A prototypical grammatical model by this definition would be the variable rule model, introduced in Labov (1969) and elaborated in Cedergren and Sankoff (1974) and Sankoff and Labov (1979): the probability of applying a rule is predicted as a weighted multiplicative combination of contextual features. Variable rule models are a subtype of logistic regression; Sankoff and Labov (1979). On this definition, then, connectionist models are also grammatical models, even if the knowledge of generalizations cannot be easily localized, since they too can be reduced to regression, e.g. Sarle (1994); see also Smolensky (1999) for a discussion of the relation between grammar and connectionism.

gorithm is the normal behavior... [O]ur brains (unlike most computers) have no need for economizing with storage space".[10] Thus, on the usage-based view of language, grammatical computation might be needed for creative use of language but not much else: as long as some structure is encountered and noticed, it can be stored and later retrieved whole, in all its morphological complexity and phonetic detail, e.g. Albright and Hayes (2003); Bybee (1985, 2001); Kapatsinski (2010c,b); Langacker (1987). In morphology, this view is supported by the common finding that the same speakers can treat known words differently from unknown words despite phonological and semantic similarity. For example, an English speaker would say that the past tense form of [gɪv] is [geɪv] and yet predict that the past tense of [kɪv] would be [kɪvd]; Albright and Hayes (2003). Kapatsinski (2010b) showed that Russian speakers spontaneously adopting English words for use on the Internet often fail to palatalize them before certain Russian suffixes (e.g., *to blog* is commonly adopted as /blogit^{j}/ rather than /bloʒit^{j}/), indicating that the palatalizing rule (g → ʒ /__i) has lost productivity. Yet, speakers always palatalize known[11] Russian words bearing the same suffixes. Assuming that the grammar is responsible for the treatment of novel words,[12] divergent treatment of a known word is a sign of the speaker having memorized how that specific word behaves.[13] The traditional conclusion is then that there are two mechanisms for production of complex forms: retrieval from the lexicon, or computation using the grammar and that retrieval usually wins over computation; Albright and Hayes (2003); Baayen (2007); Marcus et al. (1992); Pinker and

[10] Cf. Chomsky and Halle (1965):105, "a grammar should be evaluated by minimizing the total number of features specified in the lexicon and in the phonological rules... The theory of grammatical form must permit only such notations as convert considerations of generality into considerations of length... This, in fact, is the motivation for the particular decisions that have been made concerning notations in the work in generative grammar..."

[11] operationalized as 'findable in a large dictionary'

[12] As suggested by a reviewer, it is possible that sequences like [gi] and [ki] are being used to mark these words as foreign. However, I do not think this hypothesis is very plausible for this case. There are few if any borrowed words that contain these sequences, raising questions as to how an association between [gi] and [ki] and foreignness could develop. Further, the same foreign stems that are not palatalized before -i or -ik are palatalized before -ok or -ek, suggesting that the effect is specific to certain Russian suffixes, namely ones that tend to occur after consonants that are not eligible for palatalization; Kapatsinski (2010b). Finally, the reluctance to palatalize before -i is even more extreme in wug tests with novel words that are not borrowings, suggesting that the alternation has genuinely lost productivity (unpubished data).

[13] Though, on the usage-based view, it is not a prerequisite for storage; Bybee (2001):160-61.

Prince (1988).

However, constructionist approaches to grammar, exemplified by Fillmore et al. (1988); Goldberg (1995); Langacker (1987), eliminate the distinction between the grammar and the lexicon. The principal thesis of the constructionist approach is that knowledge of grammar is knowledge of constructions and the relations among them. Goldberg (1995) defines constructions as conventionalized form-meaning pairings stored in long-term memory.[14] Words are one type of construction, but larger and smaller meaningful patterns (such as phonaesthemes, morphemes, idioms, collocations, argument structure patterns, etc.) that are noticed by speakers of a language and used in production and/or perception are also constructions. All constructions are assumed to form a single system, the *constructicon* (so named on analogy with *lexicon*).

The empirical motivation for eliminating the lexicon/grammar distinction was the observation that there is a massive grey area between fixed expressions like *kick the bucket* and fully open sentence-level constructions like Subject Verb Object; Fillmore et al. (1988); Goldberg (1995). Denizens of this grey area in English include the '*Way*-Construction' *SUBJ VERB.TNS SUBJ.POSS way PP*, as in *He elbowed his way up the staircase*, and the Comparative Construction, *the X-er, the Y-er*, as in *the more, the merrier* or *the more he struggled, the faster he sank into the swamp.* In fact, *kick the bucket* itself leaves room for variability: *kicked the bucket* is an instance of the idiom, as is *kicking the proverbial bucket*, whereas *kicked a heavy bucket* and *kicked the buckets* are not. These *partially lexically specific* constructions defy a tidy division between the lexicon and the grammar.

If the lexicon/grammar distinction is eliminated, we cannot say that lexical retrieval has primacy over grammatical computation. On a constructionist approach, there is only the constructicon, usually seen as a complicated network containing hierarchies of partially redundant generalizations, e.g., *He gave her a flower* would be stored as well as *PRO give.TNS PRO NP*, and *NP V NP NP*.[15]

[14] Bybee (2001) favors a more narrow definition, where constructions are only form-meaning pairings that have open slots, thus including morphemes and larger structures but excluding phonaesthemes. Bybee and Eddington (2006) further propose that constructions are bigger than the word. The cover term for all kinds of form-meaning pairings (equivalent to Goldberg's *construction*) in Bybee's terminology would be *product-oriented schema*; for Nesset (2008), it is *first-order schema*. We adopt Goldberg's terminology here because it is simpler and more widespread.

[15] This proposal dates back at least to Langacker (1987):42, who cautioned linguists against what he called *the Rule-List Fallacy*: just because speakers induce a generalization about a set of forms does not mean that they do not *also* store the forms on whose basis the generalization is made. See also Bybee (2001):20-21 and

Under the constructionist approach, different treatment of known and unknown words is accounted for by prioritizing the most specific constructions that are compatible with the semantics that are to be expressed, e.g. Ambridge et al. (2014); Langacker (1987); Nesset (2008).[16] For example, if one wants to express the meaning GIVE.PAST, the most specific applicable construction is *gave*-GIVE.PAST, but the more general VERB_i-*ed* / ACTION_i.PAST is also applicable,[17] as might be intermediate constructions that specify some aspects of the form of the verb stem and/or the semantics of the action. To achieve the same effect that prioritizing retrieval over computation achieves in the lexicon+grammar model, one would favor *gave*-GIVE.PAST on the grounds of specificity. For a novel verb, the most specific constructions are not applicable since they do not have slots that the novel word can fit into, thus one has to fall back on a more general construction that has a compatible open slot. Storing a whole hierarchy of partially specified constructions and prioritizing the more specific applicable constructions ensures that novel words will tend to be treated like *similar* known words.[18]

In a lazy-learning approach, the priority of the specific is taken to the logical extreme. The complex hierarchies of constructions are eliminated. There are no stored generalizations, hence the priority of the

Beekhuizen et al. (2013).

[16] The same idea has also been proposed in rule-based frameworks under the names 'Paninian determinism' and the 'Elsewhere condition', e.g. Stump (2001).

[17] Construction A is more general than construction B if the features specified in A are a subset of the features specified in B. The features involved could be phonological, semantic or both.

[18] As we will argue in more detail later, this prioritization should not be absolute. The decision of which construction to apply is probabilistic, with the probability of construction selection determined by the current level of activation of that construction, in turn strongly influenced by its long-term strength. For example, in the case of the English past tense, the regular *-ed* construction is vastly stronger than the irregular constructions. If speakers always used the strongest construction applicable, irregular constructions would never apply to novel inputs. In their study of the English past tense, Albright and Hayes (2003) found that the regular output was more likely than the irregular output for every one of their novel stimuli, even ones that were very similar to existing irregulars, showing the regular construction to be dominant, in line with its high type frequency in English. Nonetheless, irregular constructions were extended to novel inputs to the extent that the novel inputs were similar to gangs of existing irregular words, and the likelihood of applying one of these irregular constructions was proportional to the statistical reliability of the construction. If construction choice were not probabilistic, the reliabilities of the weak irregular constructions would not matter, and the stronger regular construction would always be chosen. When constructions are placed in competition within a miniature artificial language, probability matching behavior is likewise observed, e.g. Kapatsinski (2010b).

most specific constructions comes for free. There is no inference during learning: speakers are not learning which features of words predict the values of other features. Novel words are treated by comparing them to similar known words stored in the lexicon (known as lexical *neighbors*). Skousen (1989) elegantly captures the insight that novel words might be treated differently from similar known words by assuming that a known word is its own closest neighbor. Skousen (1989) proposes that in order to know how to treat a word, the speaker searches the lexicon for the closest neighbor(s) of that word. Furthermore, distant neighbors influence the decision only if allowing them to weigh in on the current decision would improve the speaker's confidence in that decision. For example, if 60% of the nearest neighbors are voting for outcome 1, and 40% are voting for outcome 2, and the neighbors a little further away are 90% in favor of outcome 1, they will be allowed to influence the decision. However, if the more distant neighbors were to favor outcome 2 60% to 40%, they would not be taken into account. When the word is known, there is only one nearest neighbor (the word itself), hence more distant neighbors have no chance of influencing the decision of how the word is to be treated. In TiMBL,Daelemans and van den Bosch (2005), the same result can be achieved by weighting known words by the inverse of their distance to the word whose behavior is being predicted. If the word is actually known, its behavior will always be based on itself, as its distance from itself is zero and the inverse of zero (1/0) is positive infinity.

It is worth pointing out that all existing models of morphology and phonology that claim to be analogical, exemplar-based or lazy-learning assume segmentation into words. Words are generalizations over observed utterances, thus these models are not *completely* lazy.[19] Nonetheless, we can ask whether any further generalization is necessary or if a lexicon of words is sufficient to account for morphological and phonological creativity. I would argue that an unanalyzed lexicon is not enough: an adequate description of morphology or phonology requires task-specific weighting of sublexical features, and therefore cannot be the outcome of lazy learning.

For example, Albright and Hayes (2003) model acquisition of the

[19] Though the segmentation into words is often seen as merely a simplifying assumption, e.g. Goldinger (1998), it is not clear if the models would perform at all if the exemplars were full-fledged utterances, e.g. the performance of the analogical model of the past tense in Albright and Hayes (2003) decreases when verbs are not stripped of their prefixes, presumably because, without feature weighting, it is mislead by similarities between verbs sharing prefixes. Analogical Modeling of Language Skousen (1989) is unable to deal with more than a few features because processing costs grow exponentially as the number of features increases.

English past tense, pitting a lazy learning model against a grammatical model.[20] The past tense in English can be expressed using either the regular suffix (*-ed*, with one of the three phonologically-conditioned allomorphs, [d], [t], or [ɨd]), or one of the irregular patterns (like *drink-drank*, *think-thought*, etc.). The choice of how the past tense is expressed is influenced strongly by the phonological form of the verb stem. However, not all parts of the stem are equally informative. The identity of the final segment is much more important than the rest of the stem. For instance, if you know that the stem ends in a voiceless fricative, 351/352 times it will be affixed with *-ed* (and the [t] allomorph of *-ed* will always be chosen); if you know that the stem begins with a voiceless fricative, little can be said about the choice of the past tense expression. The importance of the stem-final segment is not just due to its overall perceptual salience: as shown by Marslen-Wilson and Tyler (1980), initial segments are more important than final ones for word recognition, since they allow the word to be recognized faster. Initial segments are also more important than final ones for picking prefixes, e.g. whether the prefix is *in-* as in *incredible* or *un-* as in *unthinkable.* The final segment is important specifically for predicting English past tense expression, since one of the exponents of past tense is a suffix. A single store of utterances that one generalizes over in a post-hoc fashion whenever any language-related task comes up would not be able to express this fact, and the lazy learning model embodying this hypothesis does in fact perform worse than the grammar-based model in Albright and Hayes (2003).

Kalyan (2012) makes the same point with respect to syntactic generalizations. He argues, based on empirical work by Ambridge and Goldberg (2008), that the acceptability of a sentence of the type *Who*$_i$ *did X verb that Y verbed* _____$_i$*?* depends on the extent to which the main clause verb foregrounds its complement clause. For example, *mumble* backgrounds the complement, and *Who did she mumble that he saw?* is of questionable acceptability. On the other hand, *say* foregrounds the complement, and *Who did she say that he saw?* is a perfectly acceptable sentence. Dąbrowska (2004) argues for an analogical account, in which the acceptability of such sentences depends on the similarity of the main clause verb to *say* and *think.* However, as Kalyan (2012):545 writes, ”how does the speaker know that in this case, similarity should

[20] Again, the important distinction for the present purposes is that the grammatical model is not lazy: Keuleers (2008) shows that Albright and Hayes (2003)'s rule-based Minimal Generalization Learner can be seen as a special case of the analogical TiMBL with particular, and probably undesirable, restrictions on feature weighting. While analogical, TiMBL does have feature weighting.

be judged with respect to foregrounding of the complement, as opposed to some other property of the verb?.. This is a problem for any exemplar model of productivity". One has to learn that foregrounding is especially important for this particular construction. One possible way to do that is to determine which features of the verbs (or indeed utterances) characterize instances of the construction / express the meaning of the construction; Goldberg (1995); Kalyan (2012); Kapatsinski (2013); see also Eddington (2004); Arndt-Lappe (2011); Daelemans et al. (2010):16 for evidence that analogical models of morphology improve with feature weighting.

Kapatsinski (2009a):Ch.4, performed a miniature artificial language experiment that is also relevant here.[21] Miniature artificial language learning is a way to empirically identify the generalizations made by human language learners on the basis of a particular linguistic experience. In this particular experiment, the learners were presented with a language in which there were two plural suffixes, *-i* and *-a*, where *-a* occurred after stems ending in [p] or [t] while *-i* occurred after stems ending in [k] or [tʃ].[22] For instance, the learners would experience that the plural of *kloup* is *kloupa* while the plural of *dretch* is *dretchi*. Importantly for the present purposes, half of the test stimuli shared everything except the final consonant with training stimuli that took a different suffix. The participants largely based their responses on the final consonant, acquiring the relationship between final consonant and suffix choice and applying the acquired knowledge to the potentially confusing test stimuli; Kapatsinski (2009a):127. If they simply memorized the training items and chose plural forms for test items by computing overall similarity between test items and training items without having learned that the final consonant is especially important, they should not have been able to perform the task accurately.[23].

MacWhinney (2001), among others, has documented transfer of first language feature weights from first to second language. The features in

[21] The aim of the experiment was not to distinguish between grammar-based and analogical models, thus this particular aspect of the design is discussed here for the first time.

[22] Languages 1 and 3 in Kapatsinski (2009a). Training consisted of auditory presentation of words paired with pictures of the referents. The referents were novel creatures. The task during training was to simply learn the words, and the singular and plural forms sharing the stem were not presented next to each other in time. The task during test was to come up with a plural form given a singular, pronouncing it aloud.

[23] An important caveat is that some participants may have anticipated that a plural-making test was coming, and therefore used a grammar-learning strategy that they would not use for language acquisition outside the lab. However, this is not a criticism that applies to the naturalistic data in Albright and Hayes (2003)

question were properties like animacy of the subject, case marking and word order, which are cues to who the agent of the action described by the sentence is. Weights of these features vary widely across languages, which can be detected by placing them in competition. For instance, does *Him hit I* mean that I hit him or that He hit me? A native Russian speaker would choose the first option while a native English speaker would choose the second. Here, the cues of word order and case marking are placed in competition. The learned weights of these cues differ in English and in Russian. Russian is a free word order language, so word order is uninformative for deciding who the agent is. Russian also has case marking on nouns, which makes case a really good cue for the identity of the agent. English is the opposite: case is relatively uninformative, since it only occurs on pronouns and is being lost even there (cf. the variation in the use of *who/whom* and *I/me*). In contrast, English word order is quite strict, thus being a very good cue to agency. MacWhinney (2001) argues that Russian speakers transfer the cue weights they learned in Russian into English. Transfer of feature weights is also well documented in phonology where learners have to acquire which acoustic cues are relevant for distinguishing words, e.g. Holt and Lotto (2006); Kondaurova and Francis (2008); Maye et al. (2008). It is difficult to see how the transfer of feature weights from first language to second language (which has an entirely different lexicon) can be accounted for in a framework where there is no long-term storage of such weights; see also Ellis (2006) for discussion.

I believe that task-specific feature weighting is part of learning a language: to acquire language, we infer which features of utterances are relevant for predicting the values of other features. Associations between feature values allow us to anticipate the predictable values in advance during word recognition; Grosjean (1980); Marslen-Wilson and Tyler (1980); Allopenna et al. (1998). They also help fill in what has been obscured by environmental and internal noise; Darcy et al. (2009); Kirov and Wilson (2013). Being able to predict something may even be intrinsically rewarding; Miller (1983); Biederman and Vessel (2006). Acquisition of sublexical associations under passive listening conditions has also been documented empirically, e.g. Aslin et al. (1998); Dell et al. (2000); Idemaru and Holt (2011).

An additional difficulty for the lazy learning approach arises from a divergence between the foundational evidence for this view in visual categorization, and studies on determinants of productivity of linguistic patterns. Nosofsky (1988) studied the categorization of simple visual stimuli, colored patches varying in brightness and saturation: the more bright and saturated examples belonged to one category while the less bright and saturated patches belonged to the other. He varied the frequency with which individual members of the categories were presented: some of the items were presented more often than others. Nosofsky found that learners were more likely to assign category membership on analogy with frequent examples than infrequent ones. This provided support for the idea that categorization is accomplished by analogy to the stored tokens of members of the category, where every token has the power to attract new category members. For colored patches, high token frequency of stimuli exemplifying a category was found to increase the attractiveness of that category for new stimuli. In contrast, studies of morphological patterns have repeatedly failed to find an advantage for patterns that are exemplified by frequent words; see Richtsmeier (2011) for a review and additional evidence. In fact, many studies suggest that high token frequency of exemplifying words makes a pattern less productive; Baayen (1992); Bybee (1995, 2001):118-126, Bybee and Brewer (1980); Eddington (2004); Hay (2001, 2003). At the same time, recognition of instances of a single word as being instances of that word is easier when the word is a frequent one; Broadbent (1967); Goldiamond and Hawkins (1958); Howes and Solomon (1951) *inter alia*. Morphological processing of known words is likewise easier if the word is frequent: Ellis and Schmidt (1997) show that it is easier to generate the past tense of a frequent verb than of an infrequent verb. Yet, it does not appear to be easier to generate the past tense of a novel verb that is similar to a frequent known verb than that of a novel verb similar to an infrequent known verb. This divergence in results for known and novel words is unexpected if categorization of new words and old words is the same process, and known words are simply their own closest neighbors.

In contrast, high token frequency of words exemplifying a pattern is expected to make the pattern less productive if patterns need to be parsed out of the exemplifying words to be extended to new words, as long as we assume that 1) words compete with their parts for recognition, and 2) this competition is affected by frequency: the more frequent a word, the easier it is to access directly, and the harder it is to access its parts; Bybee (2001):118-126, Bybee and Brewer (1980); Hay (2001); Phillips (2001). There is some empirical evidence for the claim that recognition of the same stimulus is harder when that stimulus is embedded in a high-frequency word. Healy (1976) found that *h* is harder to detect in *the* than in *thy*. Kapatsinski and Radicke (2009) found that the auditory sequence /ʌp/ is harder to detect in frequent words like /kʌp/ than in infrequent words like /pʌp/. For the colored patches of Nosofsky (1988), categorization depends on a single dimension (color) that is very easy to parse out of the stimuli. Linguistic units, like words and utterances, are highly multidimensional, making parsing out the features associated with a word class a real challenge. Nonetheless, word class extension appears to be based largely on parsing patterns out of individual words rather than on analogy with frequent words exemplifying the class.[24]

3 Grammar acquisition is a small-n, large-p problem but redundancy makes it easy

Why then are lazy learning models usually successful, despite having no feature weighting and no acquisition of feature-feature associations? The answer appears to be that language is highly redundant, in the sense that the occurrence of any given feature is predictable from a large number of other features (Hockett (1965), see Ackerman and Malouf (2013); Hayes (1999) for morphological paradigms in particular). There are many possible reasons for this state of affairs, only one of which is exemplar-based memory. The undeniable fact is that linguistic structures are highly multidimensional. The most economical descriptions of speech sounds still utilize dozens of features, e.g. Chomsky and Halle (1968). Each feature is redundantly cued by multiple acoustic cues, e.g. Wright (2004). Words usually consist of multiple speech sounds with additional suprasegmental features overlaid on top. These multidimensional structures do not fill the space of possible words evenly. The unevenness is fundamentally due to the fact that not all sequences of sublexical units are equally easy to pronounce, equally easy to perceive, and make sense in semantics. The unevenness is exacerbated because of rich-get-richer positive feedback loops operating on sublexical units: the more a morpheme, a phonaestheme, an articulatory gesture, a gestural co-ordination pattern, etc. is used, the more likely it is to be re-used in the future, e.g. Dell (1986); Martin (2007). As discussed in Yule (1944); Simon (1955); Barabási and Albert (1999) and Piantadosi (2014), such positive feedback loops produce highly skewed, Zipfian frequency distributions, making some areas of the space of possible words densely populated and some empty. Finally, as proposed by Bybee (2002), units

[24] Bybee and Eddington (2006) propose a possible counterexample, where Spanish verbs of 'becoming' are argued to be extended to new adjectives on analogy with frequent adjectives they already occur with. While Bybee and Eddington (2006) do not directly test for a frequency effect, they may be right about a special role of high-frequency words in the case of *semantic* class extension. Class extension may be less reliant on semantic features than phonological features, as suggested by the results of Gagliardi and Lidz (2014). Perhaps, word meanings are less likely to be decomposed than word forms.

used together fuse together, coming then to be re-used as an even more multidimensional chunk. As a result, any characteristic associated with a set of words is predictable from many features, simply because there are many other features that the words have in common and because features become associated with each other, forming larger constellations we call constructions.

Redundancy means that features will typically agree with each other in predicting the value of some other feature, making models that have no feature weighting perform reasonably well, e.g. Albright and Hayes (2003); Arndt-Lappe (2011); Daelemans et al. (2010); Eddington (2004). Reundancy also allows for a large degree of individual I-language variability in the presence of E-language uniformity; see also Hockett (1965); Householder (1966):99, Langacker (1987):28, Bybee (2001):32.[25] In the extreme, some speakers' heads could host exemplar models, and some could contain fairly abstract grammars, and the produced output would be essentially identical. For example, Ernestus and Baayen (2003) show that Dutch native speakers can agree whether a voiceless consonant at the end of a novel word is underlyingly voiced or voiceless, and that their judgments reflect the statistics of the lexicon. They model this behavior with two different exemplar models, stochastic Optimality Theory, classification and regression trees and a spreading activation model containing both words and sublexical features. All models perform very well, and approximately *equally* well, indicating that different speakers could learn generalizations at different levels of abstraction and still perform the task. Similarly, Divjak and Arppe (2013), while trying to predict near-synonym choice for verbs of trying in Russian, find that "quite similar levels of model fit and prediction accuracy can be achieved by selecting clearly divergent sets of properties in a model" (p.234, see fn.14 for the data) and conclude that "very different stored property combinations making up the core of the prototype [for the meaning of a verb of trying] would result in prototypes being different for every person and would make it irrelevant what learners track, as long as they track something" (see p.245,

[25] Anyone who has taken a phonology course is familiar with the fact that multiple solutions are usually possible for any given phonology problem. For instance, a phoneme inventory can often be described by a number of different feature systems, implying that it is unclear which differences among sounds are the more important ones, especially for smaller inventories. Generative linguistic theory has attempted to come up with innate constraints or procedures to predict a unique feature system for every inventory, e.g. Dresher (2009). However, Idemaru et al. (2012) document stable individual differences in which features are assumed to be distinctive by different listeners, demanding an approach that allows for distinct feature systems to co-exist within a community.

fn.15).

Redundancy also means that any reasonable model of language must be able to generate predictions for novel multidimensional structures after being trained on a small number of structures of the same type: children start talking before they can be reasonably sure what the grammar of the language is, and are able to deal with novel words and utterances, even if not in adult-like ways Berko (1958). Susrprisingly, not all models of morphology satisfy this criterion. For example, consider the rule-based model advocated by Albright and Hayes (2003), the Minimal Generalization Learner. The Minimal Generalization Learner takes in pairs of morphologically related words, splits them into change and context, and then generalizes over contexts to come up with rules. Thus rules start out very specific, particular to individual words, and only gradually become more abstract. For example, given the word pairs *bank-banked* and *link-linked*, the model would generalize the rule 0–→ed / X[-back;+syl]nk___. Suppose then that the model is presented with the verb *lick*, which it has never encountered. What would the model produce as the past tense form of this verb? The answer is that the model would have no idea: it knows of no rule applicable to this word. This is a general problem with specific-to-general learning, although it is most severe with this particular model: the system needs some way to extrapolate beyond what it has encountered to deal with inputs that are *not* minimally different from the inputs it has encountered. The problem is not apparent if a model is tested after it has acquired a highly diverse, adult-like lexicon but is very acute in modeling the early stages of language learning, e.g. as simulated by mniature artificial language learning paradigms. Possible solutions involve learning form-meaning mappings, so that the model always tries producing something that sounds like a past tense form or general-to-specific learning so that the system first becomes aware of the fact that adding -ed is a possible operation to consider (see Kapatsinski (2013) for a model combining both assumptions).

4 The problem of idiolects, and multimodel inference

There is much evidence that individuals do in fact vary in the abstractness of categories they form. In particular, low degree of abstraction appears to be correlated with autistic traits. For example, Plaisted et al. (1998) found that individuals with autism form narrower categories after exposure to a series of dot patterns. Johnson (2013b) exposed children with typical development and children with autism to a new syntactic construction (S O V-*o* = AGENT GOAL APPROACHES) and tested for generalization to instances of the same construction with novel verbs. Both groups could understand the construction when the verb had been presented in the construction during training but participants with autism were less likely to understand instances of the construction involving novel verbs. Yu (2010) found that differences in compensation for language-specific assimilation patterns were correlated with scores on a test of autistic tendencies even well below the clinical range, suggesting that the differences in degree of abstraction could be pervasive in the neurotypical population. While most work has examined perception, Mielke et al. (2013) found that both neurotypical individuals and individuals with autism spontaneously mimic voice onset times of an interlocutor in speech production but only neurotypical individuals generalize the learning to a new phoneme.

As Dąbrowska (2012) points out, individual variability in the generalizations that are formed on the basis of a particular experience with language is challenging to the traditional generative notion that linguists can describe the I-language grammar of a language shared among members of a speech community, cf. Chomsky and Halle (1965). Nonetheless, the notion of a community grammar has much to recommend itself in that it is at the level of the community that norms are enforced and conventionalization happens: the community sets the target that individuals reach for. Labov (1996):80 writes that "The central finding of sociolinguistics is that the community is the stable and systematic unit, and that the behavior of individuals cannot be interpreted without a prior knowledge of the community pattern." Importantly, however, the target set by the community is an E-language target, which can be generated internally by a number of different systems of generalizations. In reaching for this target, an individual acquires his/her own grammar, a model of the target that can re-generate the target (more or less). It is the aggregate of such individual grammars (or models) that is the community I-language grammar. It then follows that the I-language grammar of a language is the result of *multimodel inference.*[26]

Fortunately, multimodel inference methods have now become widely available in both Bayesian and frequentist frameworks. For general methods, see Burnham and Anderson (2002); Hoetting et al. (1999);

[26] In responding to Kay and McDaniel (1979)'s critique of the variationist method, which involves building a logistic regression model of linguistic behavior observed in a corpus, Sankoff and Labov (1979):201 write: "Kay and McDaniel's discussion puts far too much emphasis on the selection of a 'best' model, which was in practice never a primary consideration. On the contrary, the main use of the various models was to locate stable and robust effects that appear in all models..."

Strobl et al. (2008). For applications to linguistic data, see Baayen et al. (2013a); Barth and Kapatsinski (2014); Kuperman and Bresnan (2012); Tagliamonte and Baayen (2012). These methods involve building all plausible models of a phenomenon, and then generating predictions from the complete set of models by weighting the predictions of every individual model by how believable or predictive it is. The idea is that the prevalence of a grammar in the population of speakers of a given language variety is proportional to how good that grammar is at generating speech representative of that language variety. Conceptualizing community grammar as multimodel inference appears to nicely capture both the existence of idiolectal variation emphasized by Bloch (1948) and Dąbrowska (2012) and the relative stability of the community grammar noted by Labov (1975, 1996).[27]

5 Grammar is non-parametric

We now turn from the problem of community grammar, which I propose to be best handled using multimodel inference, to the properties of individual grammars, or models, comprising the community grammar. The first such property, discussed in the present section, is that grammar is non-parametric: the number of generalizations in the grammar (or parameters in the model) is in principle unlimited, and should grow in the course of language acquisition. This proposal is in sharp contrast to the *Principles and Parameters* approach of Chomsky (1981) and the computational models that assume this approach, e.g. Niyogi (2006); Yang (2002), and is the primary claim of the constructionist approach to grammar, as discussed in Croft (2001); Fillmore et al. (1988); Goldberg (1995).

On the constructionist approach, knowledge of grammar is knowledge of constructions and the relations among them. Crucially, just as there is no fixed universal inventory of words, *there is no fixed inventory of grammatical constructions*: different languages have different

[27] The application of multimodel inference within an individual is more controversial. For example, Baayen et al. (2013a) consider it psychologically implausible. In contrast, Beekhuizen et al. (2013):268 suggest that we should "allow for multiple (in fact, many) alternative derivations of the same sentence, with the same structure and the same semantics" within an individual, a position consistent with Langacker (1987)'s caution against the Rule/List fallacy, the assumption that an utterance can only be produced one way. We will not be able to settle this issue here. However, the idea of having many different routes to get from form to meaning is well accepted in psycholinguistics (see Baayen (2007) for a review). The existence of multiple parallel routes to get from meaning to form is likewise plausible, though is by no means a consensus position. For the multiple routes to exist, multiple analyses need to have been inferred for (parts of) the same utterance, requiring multimodel inference within an individual.

constructions, as may different speakers of the same language Croft (2001); Dąbrowska (2012). I believe this to be *the* fundamental insight of constructionist theories of grammar for statistical modeling. As pointed out by Johnson (2013a), non-parametric models are necessary for the acquisition of the lexicon. Constructionist approaches suggest that there is no fundamental distinction between the lexicon and the grammar: both words and grammatical patterns are constructions, and both are potentially infinite in number. The flexibility of non-parametric models is thus also to be harnessed for modeling the acquisition of grammatical constructions.

Typological evidence strongly indicates that there are few, if any, universal constructions. Even general patterns like $S\ O\ V$ vary in their specific range of functions across languages; Croft (2001). Furthermore, languages are not describable using a small, finite set of parameters, since every language contains constructions that seem to instantiate competing parameter settings. For instance, while English generally places determiners before nouns, there is one that can go after them, as in *exceptions galore*, thus there is no setting of the headedness parameter that describes all English constructions, or even all English determiner phrases; see Hasegawa et al. (2010).[28] If constructions are specific to a particular language, they must be learned from the input rather than genetically encoded in Universal Grammar. Since the partially lexically specific constructions are numerous (in fact, potentially infinite in number), non-parametric inference techniques are required for grammar acquisition. Construction grammarians argue that if we need to learn the huge inventory of partially lexically specific constructions, we might as well use the same mechanism to learn the more general constructions like $S\ O\ V$. Thus, the constructionist view of grammar suggests that non-parametric models are both necessary and sufficient for grammar acquisition: there is no need to posit a separate, parametric model of core grammar acquisition; Goldberg (1995).

As one would expect from a non-parametric system, the number of parameters necessary to describe the constructicon grows with language acquisition. In the most trivial sense, it is undeniable that the number of words and syntactic structures grows as more of the language is experienced. In addition, I have argued that individual constructions become more well-specified over time; Kapatsinski (2013). This is not the standard view in the constructionist literature; cf. Tomasello (2003).

[28] One does not excape the conclusion that the grammar is non-parametric if, instead of considering *galore* to be a determiner with word-specific sequencing restrictions, one instead assigns *galore* its own lexical category, as the set of lexical categories then becomes cross-linguistically variable and in principle unlimited.

However, there is, I believe, much evidence in its favor. On the meaning side, the idea of increasing specification goes back at least as far as Smoke (1932):5, who writes: ”As one learns more and more about dogs, [one's] concept of dog becomes increasingly rich, not a closer approximation to some bare element.” In subsequent work, Clark (1973); Mandler (2000), and Pauen (2002), among others, have provided empirical evidence that children's word meanings start out relatively broad, and gradually narrow over the course of development; see Rogers and McClelland (2004), for a review and computational modeling; see also Griffiths et al. (2007); Love et al. (2004) for other general-to-specific approaches to modeling categorization. On the form level, Fennell and Werker (2003) and Swingley (2007) found that children accept mispronunciations of unfamiliar words as being the same word but are less tolerant of mispronunciations of familiar words, suggesting that the form-level specification of a familiar word is more detailed. Similarly, in visual word recognition, Castles et al. (2007) showed substantial priming between minimally different spellings, e.g. *lpay* → *play*, in 3rd graders that disappeared by 5th grade, a finding they interpret as indicating increasing specificity of orthographic lexical representations. In syntax, Rowland et al. (2012) found that verb overlap between prime and target increased the amount of priming for adults and older children but not younger children, who exhibited more priming of abstract syntactic patterns independently of lexical overlap.

These findings are consistent with a view that constructions become gradually more specific over time; Kapatsinski (2013). The learner starts out ready to learn any form-meaning pairing. For example, the initial assumption (never explicit, of course) is that any kind of form can mean 'a group of multiple objects of the same kind' and that forms ending in /z/ can have any meaning whatsoever. Only gradually does the child learn that plural forms should end in /z/. Eventally, this PLURAL=...z# construction becomes so strong that it can be automatically, and counterproductively, transferred to a second language. Thus, many adult native English speakers exposed to a miniature artificial language that had plurals ending in [i] or [a] were observed to erroneously add [z] to the end of the plural forms following the vowel suffix in an elicited production task; Kapatsinski (2013). The adult speaker is no longer equally ready to learn any form-meaning pairing, as some form-meaning pairings get a boost from the speaker's prior experience with language. As discussed above, general-to-specific learning also allows the learner to deal with novel inputs that are highly dissimilar from the inputs encountered so far. Without general-to-specific learning, inability to deal with such inputs appears inevitable on a purely constructionist view, i.e. a view in which novel inputs are dealt with by selecting the construction they fit into, and would arise for the same reason that it arises with the rule-based Minimal Generalization Learner of Albright and Hayes (2003).

6 Grammar is full of complex non-crossover interactions

As noted above, a typical construction is a highly multidimensional structure that can only be fully characterized on the form level by hundreds of phonological features. Importantly, most formal features of a construction are necessary for recognizing the meaning of the construction.[29] This means that a statistical model for construction acquisition should be prepared to look for complex superadditive interactions among formal features, where a meaning of a construction can only be perceived when no formal feature of a construction is perceived to be missing.[30] The auditory signal may be missing some of the features associated with a construction (due to conventional reduction patterns, mispronunciation, acoustic noise, etc.) but the listener must believe that the speaker intended to produce that particular construction. Thus, the only deviations from the full form of the construction that can be tolerated are the ones that commonly occur and are therefore easy to undo; see Norris and McQueen (2008) for computational evidence.

For example, in the absence of noise, orthographic priming is much weaker (in both magnitude and persistence) than repetition priming. A mismatch in a single letter or phoneme appears to be sufficient to eliminate repetition priming in adults; see Castles et al. (2007), among many others, for behavioral data and Glezer et al. (2009) for evidence suggesting that specific neurons in the visual wordform area respond to specific words, firing as little when presented with one-letter-away neighbors of the words they represent as when presented with completely dissimilar words.[31] In other words, all letters in an orthographically presented

[29] While not focusing on phonological features, Langacker (1987):371 writes that a schema/construction is an "abstract characterization that is fully compatible with all the members of the category it defines" so that "membership is not a matter of degree". See also the discussion in Divjak and Arppe (2013):225-227

[30] This is presumably the intuition behind Bloomfield (1926)'s famous statement: "Such a thing as a "small difference of sound" does not exist in language." For an experienced listener in the absence of noise, a tiny acoustic difference can make the difference between a word and a non-word.

[31] Presumably one could get repetition priming for mismatched primes and targets if the mismatch could plausibly due to a common spelling or typing error (as in *langauge* for *language*, which has appeared on published covers of linguistics books)

construction are necessary to recover for long-term repetition priming to occur.[32] Importantly, words are not the only constructions in the constructicon, and repetition priming effects have been obtained for other constructions as well, including morphemes in Clahsen et al. (2003); Marslen-Wilson et al. (1994); Stanners et al. (1979), phonaesthemes in Bergen (2004), and syntactic constructions in Bock (1986); Rowland et al. (2012). Furthermore, as noted above, what constitutes an exact match changes with experience: children tolerate greater deviations from the canonical form of a word for rarely-encountered words; Swingley (2007). It appears that repetition priming can be obtained when the prime and the target do not exactly match as long as they share a construction. [33]

Note that there is much evidence that the meaning of a word can be accessed before recognition of the word is completed. For instance, Allopenna et al. (1998) shows that participants hearing words look to pictures of referents of phonologically-similar words more than they look to pictures of unrelated distractors. For example, when the listener hears 'cattle', s/he would look at a picture of a captain more than s/he would look at a picture of a doctor. The looks to competitors happen well before the presented word is completed, suggesting that the sematics of words are activated before the presented word is recognized. Ostrand et al. (2011) presented listeners with auditory words paired with videos of faces pronouncing slightly different words (e.g., auditory *pot*, visual *cot*). Listeners consciously perceived an average of the two (here, *tot*), exhibiting the well-known McGurk effect. Nonetheless, the auditorily presented word, never consciously perceived, activated its semantic associates, generating semantic priming. Revill et al. (2008) found, with an artificial lexicon, that non-motion words that sounded like motion words activated a brain area responsible for motion processing (area MT). Thus the parts of a construction are also associated with the meaning of the construction and can activate it when the con-

or there were visual noise sufficient to believe that one has misperceived the letter but I am not aware of any work on this question.

[32] As Armstrong et al. (1983) argue, features can be necessary despite instances of their values being difficult to identify. For instance, to me a stool cannot have a back that is designed to lean against. The absence of such a back is, to me, a necessary feature of a stool. However, I may not recognize whether a given instance of a back is made to lean against, thus my stool identification procedure is noisy; see also Wierzbicka (1990).

[33] Bybee (2001), and Bybee and Moder (1983) offer a contrasting view, in which constructions do not have necessary features. However, Albright and Hayes (2003) show that the same data can be captured without abandoning necessary features, as long as partially redundant constructions are allowed (which they are in Bybee's model).

struction is being perceived. Nonetheless, something special appears to happen when all the features are perceived: the construction is consciously recognized, and its activation obtains strength and longevity. This is one sense in which feature interaction in word recognition is superadditive: the parts of a construction are associated with its meaning but the whole is more than the sum of its parts.

In addition, van den Bosch and Daelemans (2013):312-316 examined the similarity spaces of examples that can be used to subserve grammatical generalization for a variety of prediction tasks, including prediction of the plural forms of German nouns, diminutive forms of Dutch nouns. and prepositional phrase attachment for English sentences. They looked for regions in the similarity space in which all examples behaved consistently with respect to the task. Such regions were found to contain on average only 6-13 types, leading van den Bosch and Daelemans (2013):314 to conclude that "the example spaces of these tasks are highly disjunct with respect to the clusteredness of examples mapping to the same outcome". A construction describing the examples within such a uniform region would therefore usually be quite specific: for a word to fit into a well-circumscribed region of the similarity space, it must have a specific set of individually necessary and jointly sufficient features realized in the right order.[34] It is not clear how models that do not allow for feature interactions, e.g. the Naive Discriminative Learner of Baayen et al. (2011, 2013b) could account for such data (see also Minsky and Papert (1969) for the same criticism of an earlier generation of two-layer perceptrons).[35]

[34] van den Bosch and Daelemans (2013) use this finding as a motivation for lazy learning: abductive inference on the basis of nearest examples would describe such a disjoint space very well. However, while sublexical constructions tend to be fairly specific *on average*, much more general ones are also found. For example, Albright and Hayes (2003) document that almost all verbs ending in a voiceless fricative in English take the regular *-ed* past tense, and that novel forms that end in a voiceless fricative are very likely to do so as well. This construction subsumes hundreds of English verbs. The type frequency distribution of constructions may be expected to be highly skewed because of a rich-get-richer dynamic in construction use: the more a construction is used, the more likely it should come to be re-used and to acquire new instantiating words or expressions. Therefore, average construction type frequency may greatly underestimate how general constructions can get.

[35] This is acknowledged by Baayen et al. (2013b):341, who write: "We note here that it is conceivable that many n-grams have their own semantic idiosyncrasies, just as many derived words and compounds have meanings that are not strictly decompositional. Any n-gram with an idiosyncratic sense will require an independent meaning outcome in our model. Without sense annotations allowing us to distinguish between non-decompositional and decompositional n-grams, the modeling of the finer semantic details of word n-grams is currently not possible." However, this admission may not be going far enough. Bybee (2001):160 points out that phono-

Classification and regression trees are one possible inference model that is designed to look for complex non-crossover interactions; see Labov (1969); Daelemans et al. (1999); Ernestus and Baayen (2003); Baayen et al. (2013a), and Kapatsinski (2013) for linguistic applications. In Kapatsinski (2013), classification and regression trees are applied to the problem of acquiring sublexical constructions, phonological structures associated with a certain cell in a morphological paradigm. I use phonological features of words as predictors of whether a word occurs in, say, the sublexicon of plural nouns. The tree finds the most predictive feature and places it on top, and then adds extra features to the extent that they help predict occurrence in the set of plural forms, given all the features already in the tree. I show that sublexical plural constructions can then be read off this tree: they are the paths that either end in leaves describing existing plural nouns, or non-terminal nodes that dominate such leaves. For example, PLURAL=...Vtʃi#, PLURAL=...tʃi# and PLURAL=...i# (where # is a word boundary) is a hierarchy of constructions extracted for a language that has a plural suffix *-i*, which often follows [tʃ]-final stems, in which the [tʃ] is usually preceded by a vowel. As the language is acquired, the construction hierarchy grows, more specific constructions being added on top of more general constructions, starting with PLURAL=...i#, then adding PLURAL=...tʃi#, and finally PLURAL=...Vtʃi#.[36]

7 What is in the grammar: Constructions+

While Goldberg (2002) proposed that constructions are all there is to grammar, certain phenomena in phonology and morphology appear to militate against this view. In particular, additional machinery appears to be required for the acquisition of non-lexical phonology and of arbitrary paradigmatic mappings. Phonological knowledge appears to include knowledge of sequencing patterns (*phonotactics*). Importantly,

logical and semantic change specific to individual words or phrases, which results in their loss of compositionality could not be word- or phrase-specific ”if there were not already material stored there on which to register the changes. That is, the vowel in *I don’t know* could not reduce to schwa *in this particular phrase* unless the phrase were present in storage. Similarly, a new discourse function could not be assigned *to this phrase* unless it was already present as an autonomous unit. Thus, both the functional and phonological changes attest to the *prior* autonomy of these phrases...” (emphasis mine)

[36] I assume that constructions compete with a tendency to repeat the known form (in this case, singular). As a result, early in the acquisition of the language, participants may simply add *-i* to a stem like [bluk], producing [bluki]. Once PLURAL=...tʃi# becomes strong enough, they might produce [bluktʃi]. Finally, once PLURAL=...Vtʃi# is strong enough, [blutʃi] will be produced, see Kapatsinski (2013) for details.

phonotactic knowledge can be learned without learning anything about word meanings, purely from the statistics of a meaningless stream of sounds: Aslin et al. (1998), or experience with pronouncing meaningless non-words: Dell et al. (2000). It is difficult to see how these results could be accounted for if all of grammar acquisition consisted of learning meaning-linked constructions. Purely form-level sublexical categories and associations between these categories appear to be necessary.

It is important to point out that the existence of pure phonology does not mean that there is an architectural restriction such that phonological units cannot become associated with meanings unless they are combinable in syntax, contra generative views like Chomsky (1981, 1993). The psychological reality of phonaesthemes (like *gl-* = LIGHT in *glow*, *glint*, etc.), as documented in Bergen (2004), strongly suggests that language-specific sound-meaning associations can be acquired even for sounds that do not enter into combinations with other units. Baayen et al. (2011, 2013b)'s success in modeling a variety of phenomena in word recognition using only direct associations between letter unigrams and bigrams on the one hand and semantic features on the other likewise suggests that such an architectural distinction is unprofitable. The point I wish to argue here is simply that not all linguistic units are extracted because they are predictive of meanings. Some may instead be used simply to predict other units at the form level, or to deal with variation in pronunciation.

In morphology, constructions have difficulty with accounting for the ability of speakers to acquire arbitrary paradigmatic mappings, documented by Becker and Gouskova (2012); Booij (2010); Nesset (2008) and Pierrehumbert (2006). Such mappings are important for deriving new forms of known words. For example, a Russian speaker knows that the genitive plural of a novel pseudoword *flarnikrap* would be *flarnikrapov* while the genitive plural of *flarnikrapa* would be *flarnikrap*. This set of mappings (*0-ov*, *a-0*) is phonetically arbitrary and must be learned. The paradigmatic pairings between the suffixes are not captured by a grammar that contains only form-meaning associations. Paradigmatic form-form or construction-construction associations appear to also be necessary.

Experimental work suggests that arbitrary paradigmatic mappings are much more difficult to learn compared to form-meaning pairings, or constructions, e.g. Frigo and MacDonald (1998). This is not a priori surprising in that acquisition of such mappings requires comparison between two constructions that do not commonly co-occur, a highly demanding task.[37] However, Ackerman and Malouf (2013) and Hayes (1999) suggest that morphological paradigms in natural languages appear to have a very high degree of paradigmatic redundancy, such that any form in a paradigm is predictable from any other form, which may help the learnability of such systems. It remains an open question whether paradigmatic associations are always mappings between constructions, as proposed by Nesset (2008) and Kapatsinski (2013), cf. Ackerman and Malouf (2013). If they are, then constructions could be argued to be a developmental pre-requisite for paradigms, and models of paradigm learning could be built on top of models of construction learning.

8 Grammar acquisition is softly biased

As was pointed out by Mitchell (1980), all learners are biased. For example, every set of positive examples of category members is consistent with two extreme hypotheses: only the experienced examples are in the category, or everything is in the category. Real learners fall somewhere between the two extremes. Category breadth biases of this kind have long been examined in the literature on concept acquisition, e.g. Rogers and McClelland (2004); Xu and Tenenbaum (2007), and are beginning to be examined in other domains of linguistics as well, e.g. Johnson (2013b); Dąbrowska and Szczerbiński (2006); Kapatsinski et al. (2013); Yu (2010). In addition to biases that have to do with category breadth, there appear to be biases against stem changes Kapatsinski (2013); Zuraw (2000), especially major ones: White (2014); Stave et al. (2013), biases against interactions between non-shared fea-

[37] For example, in visual perception, Mitroff et al. (2004) argue, based on evidence from change blindness experiments, that "nothing compares two views". Something does appear to compare two "views" in language learning, else purely paradigmatic mappings would be unlearnable. However, form comparison is not as easy as exclusively rule-based models, such as Albright and Hayes (2003); Chomsky and Halle (1965); Reiss (2004), which cannot learn *anything* without making a between-form comparison, would lead us to believe. In particular, Kapatsinski (2012, 2013) shows that giving language learners examples like SG=[blutʃ] / PL=[blutʃi] increases the likelihood that they will think that the plural of a singular like [slaɪt] is [slaɪtʃi], rather than [slaɪti]. This is unexpected if paradigmatic mappings are acquired exclusively on the basis of form comparisons: the relationship between SG=[blutʃ] / PL=[blutʃi] is the same as the relationship between SG=[slaɪt] / PL=[slaɪti]: 0 → i, whereas the relationship between SG=[slaɪt] / PL=[slaɪtʃi] is different (t → tʃi). On the other hand, the results are expected if participants are learning generalizations over single forms rather than form pairs, constructions like PL=...tʃi#. Frigo and MacDonald (1998) is part of a long line of studies trying to teach participants arbitrary paradigmatic mappings, such as 'if SG=...i# then PL=...de#, while if SG=...u# then PL=...la#', which have met with very limited success in the absence of additional within-form cues as to which suffix is appropriate.

tures of consonants and vowels: Becker et al. (2011); Moreton (2008a), and biases against category structures involving cross-over interactions among features that do not form perceptual units: Kapatsinski (2009b); Pycha et al. (2003).

It is, of course, impossible to show that something is impossible to acquire, as such a demonstration would require presenting the learners with infinite data. Furthermore, for many of the attested biases, we know that the bias is a soft one: it can be overridden with enough learning data, e.g. Moreton (2008a); Schane et al. (1975); Wilson (2006). Patterns that are difficult to acquire in the laboratory can nonetheless be productive in at least a minority of natural languages, indicating that they *can* be learned given enough input, and enough input of the right kind. For instance, Stave et al. (2013) find that labial palatalization appears to be harder to learn than coronal or velar palatalization (p → tʃ vs. k → tʃ or t → tʃ) before -a. Yet, Ohala (1978) notes that Southern Bantu has labial palatalization in the absence of coronal or velar palatalization.[38] Purely paradigmatic mappings appear to be hard to acquire, e.g. Frigo and MacDonald (1998). However, they do appear to be learned in the course of natural language acquisition: Becker and Gouskova (2012); Pierrehumbert (2006). Thus, the acquisition biases against large stem changes and arbitrary paradigmatic mappings robustly observed in language learning experiments must be soft biases.

Bayesian models provide a way to capture soft biases in a principled manner, e.g. Johnson (2013a); Moreton (2008b); Xu and Tenenbaum (2007). However, it is often unclear whether a certain bias is properly thought of as being a property of the inferential process (the prior being actively used by the learner), or an outcome of how the data are experienced by the learner due to noise in the environment and imperfections of human perception, memory and motor control (*inductive bias* vs. *channel bias* in Moreton's terminology). Biases that come from prior experience, such as the transfer of cue weights from L1 to L2, may be especially good candidates for learner-internal influences (inductive bias). Biases that come from the biology of peripheral motor and sensory systems seem to influence the experienced data and the motor output rather than the inference process. The latter kind of bias is, perhaps, more fruitfully handled by embedding the inferential system within a larger system of interacting embodied agents, e.g. Cangelosi and Riga (2006). See Moreton (2008a,b) vs. Kapatsinski (2011); Xu and Tenenbaum (2007) vs. Spencer et al. (2011) for debates on the loci

[38] See Anderson (1981); Bach and Harms (1972); Blevins (2004); Hayes et al. (2009) for many additional examples.

of documented biases.

9 Grammar application is stochastic

Beginning with Labov (1969), grammatical theory has gradually come to terms with the fact that grammar application is probabilistic (see Coetzee and Pater (2011) for a review). This tendency is so ubiquitous that Hayes et al. (2009):826 call it a Law. They formulate it as "Speakers of languages with variable lexical patterns respond stochastically when tested on such patterns. Their responses aggregately match the lexical frequencies". The frequencies in question are type frequencies, and not token frequencies: the number of distinct words exemplifying a pattern is reflected in the probability of a novel word exemplifying the pattern. In other words, language learning involves *probability matching*. For example, Kapatsinski (2010b) exposed English speakers to a language in which 70% of the nouns took the plural -*i*, and 30% took the plural -*a*. When presented with a new noun, the learners pluralized the new noun with -*i* about 70% of the time and with -*a* about 30% of the time.

Probability matching is not specific to grammar, or even to the human species. For example, a cockroach, when shocked 30% of the time in one arm of a T-maze and 70% of the time in the other, would pick the arm where shocks are less likely 70% of the time Longo (1964). Despite its ubiquity, the behavior remains puzzling, in that it does not maximize the probability of being correct. Thus, if the cockroach always picked the arm of the maze that is less likely to deliver a shock, it would be shocked .3*1+.7*0=30% of the time. Probability matching results in the cockroach being shocked .3*.7+.7*.3=42% of the time. By doing probability matching, the cockroach fails to minimize its probability of experiencing an electric shock. Likewise, in language learning, using -*i* 70% of the time to pluralize the noun does not maximize the probability of picking the correct plural, as pointed out by Kay and McDaniel (1979):156. However, selection of the less likely pattern of behavior for production can, perhaps, be justified on the grounds of the need for practice to maintain the pattern in one's repertoire (lest one gets too stuck in one's ways in an environment where the future is uncertain, and the causes for the observed variation are unknown to the learner). It might also be explained by the greater salience of rare events compared to common ones (the shocks might be more painful when they occur in the safer arm of the maze and are relatively unexpected; the occurrences of the rarer linguistic pattern might be more surprising and therefore more noticeable). Whatever its cause, probability matching appears to be a robust phenomenon in linguistic generalization.[39]

10 Summary and conclusion

In this paper, I have argued that grammar involves abstraction, driven by the need for prediction, and that abstraction involves statistical inference. This inference process is biased, but the biases are weak enough that parametric models of grammar are untenable. Non-parametric techniques are necessary to model the resulting system. The soft biases need to be incorporated into the acquisition model, both in the form of Bayesian priors on models and in the form of limitations on perception, memory and motor control. Finally, I have argued that language is redundant, in the sense that any feature in an utterance is predictable from many others. This redundancy allows individuals with very different mental representations of language to speak essentially alike, obeying the norms of the speech community. I have argued that uncovering the set of individual grammars underlying the linguistic behavior of a community, as represented by a corpus, requires multimodel inference techniques. Finally, I have argued that grammar involves complex non-crossover interactions among weighted features, where the whole is often greater than the sum of its parts but the parts are nonetheless individually associated with the same outcome, and that the complexity of the learned interactions grows with experience.

I believe that these general principles are consistent with much work in the usage-based constructionist approach, and hope that they may be useful for future development of computational models of grammatical knowledge. There are models that are consistent with many of these principles. Of particular note, perhaps, are Bayesian non-parametrics reviewed in Johnson (2013a); random forests of conditional inference

[39] Empirical studies have documented deviations from probability matching, though such cases seem to involve situations where one pattern is overwhelmingly dominant. These deviations are sometimes in the direction of regularization, the more dominant response is selected 100% of the time: Ferdinand et al. (2013); Kam and Newport (2005) but sometimes in the direction of random guessing, where the less dominant response being selected more often than expected: Lindskog et al. (2013). It is not yet clear what accounts for these discrepancies. Possible explanations for regularization include inductive bias, or encoding failures, where the less dominant pattern can simply be missed if none of its occurrences are noticed; see Ferdinand et al. (2013); Kam and Newport (2005); Perfors (2011) for discussion. Deviations in the direction of random guessing might perhaps be attributed to the participants being cautious in inferring the existence of a frequency difference as hypothesized by Albright and Hayes (2003), or the phenomenon of habituation, whereby one gets bored with a frequently-experienced stimulus, and pays more attention to the more novel, and hence more surprising stimuli; see Harris (1943); Thompson (2009) for reviews.

trees, Strobl et al. (2008), discussed in Barth and Kapatsinski (2014), as well as in Tagliamonte and Baayen (2012) and Baayen et al. (2013a); non-parametric models with probabilistic generative representations like Albright and Hayes (2003); Hayes and Wilson (2008), and Labov (1969); analogical models with feature weighting, as in Daelemans and van den Bosch (2005); Nosofsky (1986); and neuroconstructivist multilayer perceptrons developed by Westermann and Ruh (2012). However, there is no model that is consistent with *all* of the above principles. To the extent that the principles are convincing, much work remains to be done.

References

Ackerman, F. and R. Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89:429–464.

Albright, A. and B. Hayes. 2003. Rules vs. analogy in English past tenses: A computational / experimental study. *Cognition* 90:119–161.

Allopenna, P. D., J. S. Magnuson, and M. K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language* 38:419–439.

Ambridge, B. and A. E. Goldberg. 2008. The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics* 19:357–389.

Ambridge, B., J. M. Pine, C. F. Rowland, C. F. Freudenthal, and F. Chang. 2014. Avoiding dative overgeneralization errors: semantics, statistics or both? *Language, Cognition and Neuroscience* 29:218–243.

Anderson, S. R. 1981. Why phonology isn't "natural". *Linguistic Inquiry* 12:493–539.

Armstrong, S. L., L. R. Gleitman, and H. Gleitman. 1983. What some concepts might not be. *Cognition* 13:263–308.

Arndt-Lappe, S. 2011. Towards an exemplar-based model of stress in English noun-noun compounds. *Journal of Linguistics* 47:549–585.

Aslin, R. N., J. R. Saffran, and E. L. Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9:321–324.

Baayen, R. H. 1992. Quantitative aspects of morphological productivity. In G. Booij and J. van Marle, eds., *Yearbook of Morphology 1991*, pages 109–149. Kluwer.

Baayen, R. H. 2007. Storage and computation in the mental lexicon. In G. Jarema and G. Libben, eds., *The mental lexicon: Core perspectives*. Elsevier.

Baayen, R. H., A. Endresen, L. A. Janda, A. Makarova, and T. Nesset. 2013a. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37:253–291.

Baayen, R. H., P. Hendrix, and M. Ramscar. 2013b. Sidestepping the combinatorial explosion: An explanation of *n*-gram frequency effects based on Naive Discriminative Learning. *Language and Speech* 56:329–347.

Baayen, R. H., P. Milin, D. Filipovic Durdevic, P. Hendrix, and M. Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118:438–482.

Bach, E. and R. T. Harms. 1972. How do languages get crazy rules? In R. Stockwell and R. Macaulay, eds., *Linguistic change and generative theory*, pages 1–21. Indiana University Linguistics Club.

Baese-Berk, M. and M. Goldrick. 2009. Mechanisms of interaction in speech production. *Language and Cognitive Processes* 24:527–554.

Barabási, A. and R. Albert. 1999. The emergence of scaling in complex networks. *Science* 286:509–512.

Barth, D. and V. Kapatsinski. 2014. A multimodel inference approach to categorical variant choice: Construction, priming and frequency effects on the choice between full and contracted forms of *am*, *are* and *is*. University of Oregon.

Becker, M. and M. Gouskova. 2012. Source-oriented generalizations as grammar inference in Russian vowel deletion. New York University and Indiana University.

Becker, M., N. Ketrez, and A. Nevins. 2011. The Surfeit of the Stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87:84–125.

Beekhuizen, B., R. Bod, and W. Zuidema. 2013. Three design principles of language: The search for parsimony in redundancy. *Language and Speech* 56:265–290.

Bergen, B. 2004. The psychological reality of phonaesthemes. *Language* 80:290–311.

Berko, J. 1958. The child's learning of English morphology. *Word* 14:150–177.

Biederman, I. and E. A. Vessel. 2006. Perceptual pleasure and the brain. *American Scientist* 94:249–255.

Blevins, J. 2004. *Evolutionary Phonology*. Cambridge University Press.

Bloch, B. 1948. A set of postulates for phonetic analysis. *Language* 24:3–46.

Bloomfield, L. 1926. A set of postulates for the science of language. *Language* 2:153–164.

Bock, J. K. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18:355–387.

Booij, G. 2010. *Construction Morphology*. Oxford University Press.

Broadbent, D. E. 1967. Word-frequency effect and response bias. *Psychological Review* 74:1–15.

Browman, C. P. and L. Goldstein. 1992. Articulatory Phonology: An overview. *Phonetica* 49:155–180.

Burnham, K. P. and D. R. Anderson. 2002. *Model selection and multimodel inference: A practical Information-Theoretic approach, 2nd edition*. Springer.

Bybee, J. 1985. *Morphology: A study of the relation between meaning and form*. John Benjamins.

Bybee, J. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10:425–455.

Bybee, J. 2001. *Phonology and language use*. Cambridge University Press.

Bybee, J. 2002. Sequentiality as the basis of constituent structure. In T. Givón and B. F. Malle, eds., *The evolution of language out of prelanguage*, pages 109–134. John Benjamins.

Bybee, J. 2003. Cognitive processes in grammaticalization. In M. Tomasello, ed., *The new psychology of language*, vol. 2. Lawrence Erlbaum.

Bybee, J. 2006. From usage to grammar: The mind's response to repetition. *Language* 82:711–733.

Bybee, J. and M. A. Brewer. 1980. Explanation in morphophonemics: Changes in Provençal and Spanish preterite forms. *Lingua* 52:201–242.

Bybee, J. and D. Eddington. 2006. A usage-based approach to spanish verbs of 'becoming'. *Language* 82:323–355.

Bybee, J. and C. Moder. 1983. Morphological classes as natural categories. *Language* 59:265–289.

Cangelosi, A. and T. Riga. 2006. An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science* 30:673–689.

Castles, A., C. Davis, P. Cavalot, and K. Forster. 2007. Tracking the acquisition of orthographic skills in developing readers: Masked priming effects. *Journal of Experimental Child Psychology* 97:165–182.

Cedergren, H. J. and D. Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50:333–355.

Chomsky, N. 1975. *The logical structure of linguistic theory*. Plenum Press.

Chomsky, N. 1981. *Lectures on Government and Binding*. Foris.

Chomsky, N. 1986. *Knowledge of language: Its nature, origins and use*. Praeger.

Chomsky, N. 1993. A minimalist program for linguistic theory. In K. Hale and S. J. Keyser, eds., *The view from Building 20*, pages 1–52. MIT Press.

Chomsky, N. and M. Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics* 1:97–138.

Chomsky, N. and M. Halle. 1968. *The sound pattern of English*. Harper and Row.

Clahsen, H., I. Sonnenstuhl, and J. P. Blevins. 2003. Derivational morphology in the German mental lexicon: A Dual Mechanism account. In R. H. Baayen and R. Schroeder, eds., *Morphological structure in language processing*. Mouton de Gruyter.

Clark, E. V. 1973. What's in a word? on the child's acquisition of semantics in his first language. In T. E. Moore, ed., *Cognitive development and the acquisition of language*. Academic Press.

Coetzee, A. W. and J. Pater. 2011. The place of variation in phonological theory. In J. Goldsmith, J. Riggle, and A. C. L. Yu, eds., *The handbook of phonological theory*. Wiley-Blackwell.

Croft, W. 2001. *Radical Construction Grammar: Syntactic theory in typological perspective*. Oxford University Press.

Daelemans, W. and A. van den Bosch. 2005. *Memory-based language processing*. Cambridge University Press.

Daelemans, W., A. van den Bosch, and J. Zavrel. 1999. Forgetting exceptions is harmful in language learning. *Machine Learning* 34:11–43.

Daelemans, W., J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2010. Timbl: Tilburg memory based learner, version 6.3, reference guide. Tech. Rep. 10-01, ILK Research Group, Tilburg University, Tilburg.

Darcy, I., F. Ramus, A. Christophe, K. Kinzler, and E. Dupoux. 2009. Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, and R. van de Vijver, eds., *Variation and gradience in phonology*. Mouton de Gruyter.

Deacon, T. 1997. *The symbolic species: The co-evolution of language and the human brain*. Penguin Press.

Dell, G. S. 1986. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93:283–321.

Dell, G. S., K. D. Reed, D. R. Adams, and A. S. Meyer. 2000. Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26:1355–1367.

Divjak, D. and A. Arppe. 2013. Extracting prototypes from exemplars. what can corpus data tell us about concept representation? *Cognitive Linguistics* 24:221–274.

Dresher, B. E. 2009. *The contrastive hierarchy in phonology*. Cambridge University Press.

Dąbrowska, E. 2004. *Language, mind and brain: Some psychological and neurological constraints on theories of grammar*. Edinburgh University Press.

Dąbrowska, E. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2:219–253.

Dąbrowska, E. and M. Szczerbiński. 2006. Polish children's productivity with case marking: the role of regularity, type frequency, and phonological coherence. *Journal of Child Language* 33:559–597.

Eddington, D. 2000. Spanish stress assignment within Analogical Modeling of Language. *Language* 76:92–109.

Eddington, D. 2004. Issues in modeling language processing analogically. *Lingua* 114:849–871.

Ellis, N. C. 2006. Selective attention and transfer phenomena in l2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics* 27:164–194.

Ellis, N. C. and R. Schmidt. 1997. Rules or associations in the acquisition of morphology? the frequency by regularity interaction in human and PDP learning of morphosyntax. *Language and Cognitive Processes* 13:307–336.

Ernestus, M. and R. H. Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79:5–38.

Fennell, C. T. and J. F. Werker. 2003. Early word learners' ability to access phonetic detail in well-known words. *Language and Speech* 46:254–264.

Ferdinand, V., B. Thompson, S. Kirby, and K. Smith. 2013. Regularization behavior in a non-linguistic task. *Proceedings of the Annual Meeting of the Cognitive Science Society* 35:436–441.

Fillmore, C. J., P. Kay, and M. C. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of 'let alone'. *Language* 64:501–538.

Frigo, L. and J. MacDonald. 1998. Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language* 39:218–245.

Gagliardi, A. and J. Lidz. 2014. Statistical insensitivity in the acquisition of Tsez noun classes. *Language* 90:58–89.

Glezer, L. S., X. Jiang, and M. Reisenhuber. 2009. Evidence for highly selective neuronal tuning to whole words in the "visual word form area". *Neuron* 62:199–204.

Goldberg, A. E. 1995. *Constructions: A Construction Grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. 2002. Surface generalizations: An alternative to alternations. *Cognitive Linguistics* 13:327–356.

Goldiamond, I. and W. F. Hawkins. 1958. Vexierversuch: the log relationship between word-frequency and recognition obtained in the absence of stimulus words. *Journal of Experimental Psychology* 56:457–463.

Goldinger, S. D. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological Review* 105:251–279.

Goldsmith, J. To appear. Towards a new empiricism for linguistics. In A. Clark, N. Chater, and A. Perfors, eds., *Empiricist approaches to language learning*.

Griffiths, T. L., K. R. Canini, A. N. Sanborn, and D. J. Navarro. 2007. Unifying rational models of categorization via the hierarchical dirichlet process. *Proceedings of the Annual Meeting of the Cognitive Science Society* 29:323–328.

Grosjean, F. 1980. Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics* 28:267–283.

Harris, J. D. 1943. Habituatory response decrement in the intact organism. *Psychological Bulletin* 40:385422.

Hasegawa, Y., R. Lee-Goldman, K. H. Ohara, S. Fujii, and C. J. Fillmore. 2010. On expressing measurement and comparison in English and Japanese. In H. C. Boas, ed., *Contrastive studies in Construction Grammar*, pages 169–200. John Benjamins.

Hauser, M., N. Chomsky, and T. Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298:1569–1579.

Hay, J. 2001. Relative frequency in morphology: Is everything relative? *Linguistics* 39:1041–1070.

Hay, J. 2003. *Causes and consequences of word structure*. Routledge.

Hayes, B. 1999. Phonological restructuring in yidin and its theoretical consequences. In J. Goldsmith, J. Riggle, and A. C. L. Yu, eds., *The derivational residue in phonology*. John Benjamins.

Hayes, B. and C. Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Hayes, B., K. Zuraw, P. Siptár, and Z. Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85:822–863.

Healy, A. F. 1976. Detection errors on the word ”the”: Evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception and Performance* 2:235–242.

Hintzman, D. L. 1986. ”schema abstraction” in a multiple-trace memory model. *Psychological Review* 93:411–428.

Hockett, C. F. 1965. Sound change. *Language* 41:185–204.

Hoetting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging. *Statistical Science* 14:382–401.

Holt, L. L. and A. J. Lotto. 2006. Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America* 119:3059–3071.

Hooper, J. Bybee. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie, ed., *Current progress in historical linguistics*, pages 96–105. North-Holland.

Householder, F. W. 1966. Phonological theory: A brief comment. *Journal of Linguistics* 2:99–100.

Howes, D. H. and R. L. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology* 41:401–410.

Idemaru, K. and L. L. Holt. 2011. Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance* 37:1939–1956.

Idemaru, K., L. L. Holt, and H. Seltman. 2012. Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *Journal of the Acoustical Society of America* 132:3950–3962.

Johnson, M. 2013a. Language acquisition as statistical inference. In S. R. Anderson, J. Moeschler, and F. Reboul, eds., *The language-cognition interface*, pages 109–134. Droz.

Johnson, M. A. 2013b. *The cognitive and neural basis of language learning: Investigations in typical and autistic populations*. Phd dissertation, Princeton University.

Kalyan, S. 2012. Similarity in linguistic categorization: The importance of necessary properties. *Cognitive Linguistics* 23:539–554.

Kam, C. L. Hudson and E. Newport. 2005. Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development* 1:151–195.

Kapatsinski, V. 2009a. *The architecture of grammar in artificial grammar learning: Formal biases in the acquisition of morphophonology and the nature of the learning task*. Phd dissertation, Indiana University.

Kapatsinski, V. 2009b. Testing theories of linguistic constituency with configural learning: The case of the English syllable. *Language* 85:248–277.

Kapatsinski, V. 2010a. Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech* 53:71–105.

Kapatsinski, V. 2010b. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory Phonology* 1:361–393.

Kapatsinski, V. 2010c. What is it i am writing? lexical frequency effects in spelling Russian prefixes: Uncertainty and competition in an apparently regular system. *Corpus Linguistics and Linguistic Theory* 6:157–215.

Kapatsinski, V. 2011. Modularity in the channel: The link between separability of features and learnability of dependencies between them. *Proceedings of the International Congress of Phonetic Sciences* 17:1022–1025.

Kapatsinski, V. 2012. Which statistics do learners track? rules, constraints, or schemas in (artificial) language learning. In S. T. Gries and D. Divjak, eds., *Frequency effects in language learning and processing*, pages 53–82. Mouton de Gruyter.

Kapatsinski, V. 2013. Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language* 89:110–148.

Kapatsinski, V., P. Olejarczuk, and M. A. Redford. 2013. Perceptual learning of intonation contours: Adult are more narrow-minded than children. University of Oregon.

Kapatsinski, V. and J. Radicke. 2009. Frequency and the emergence of prefabs: Evidence from monitoring. In R. Corrigan, E. Moravcsik, H. Ouali, and K. Wheatley, eds., *Formulaic Language. Vol. II: Acquisition, loss, psychological reality, functional explanations*, pages 499–522. John Benjamins.

Kay, P. and C. McDaniel. 1979. On the logic of variable rules. *Language in Society* 8:151–187.

Keuleers, E. 2008. *Memory-based learning of inflectional morphology*. Phd dissertation, University of Antwerp.

Kirov, C. and C. Wilson. 2013. Bayesian speech production: Evidence from latency and hyperarticulation. *Proceedings of the Annual Meeting of the Cognitive Science Society* 35:788–793.

Kondaurova, M. V. and A. L. Francis. 2008. The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *Journal of the Acoustical Society of America* 124:3959–3971.

Kuperman, V. and J. Bresnan. 2012. The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66:588–611.

Labov, W. 1969. Deletion, contraction, and inherent variability of the English copula. *Language* 45:715–762.

Labov, W. 1975. *What is a linguistic fact?*. Peter de Ridder Press.

Labov, W. 1996. When intuitions fail. *Chicago Linguistic Society* 32:77–106.

Labov, W. 2001. *Principles of language change: Social factors*. Blackwell.

Langacker, R. W. 1987. *Foundations of Cognitive Grammar. Vol. 1: Theoretical prerequisites*. Stanford University Press.

Liljencrants, J. and B. Lindblom. 1972. Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* 48:839–862.

Lindskog, M., A. Winman, and P. Juslin. 2013. Is it time Bayes went fishing? Bayesian probabilistic reasoning in a category learning task. *Proceedings of the Annual Meeting of the Cognitive Science Society* 35:906–911.

Longo, N. 1964. Probability-learning and habit-reversal in the cockroach. *The American Journal of Psychology* 77:29–41.

Love, B. C., D. L. Medin, and T. M. Gureckis. 2004. Sustain: A network model of category learning. *Psychological Review* 111:309–322.

MacWhinney, B. 2001. The Competition Model: The input, the context, and the brain. In P. Robinson, ed., *Cognition and second language instruction*, pages 69–90. Cambridge University Press.

Mandler, J. M. 2000. Perceptual and conceptual processes in infancy. *Journal of Cognition and Development* 1:3–36.

Marcus, G. F., S. Pinker, M. Ullman, M. Hollander, T. J. Rosen, and F. Xu. 1992. *Overregularization in language acquisition*, vol. 57 of *Monographs of the Society for Child Language Development*.

Marslen-Wilson, W. and L. Tyler. 1980. The temporal structure of spoken language understanding. *Cognition* 8:1–71.

Marslen-Wilson, W., L. Tyler, R. Waksler, and L. Older. 1994. Morphology and meaning in the English mental lexicon. *Psychological Review* 101:3–33.

Martin, A. T. 2007. *The evolving lexicon*. Phd dissertation, UCLA.

Maye, J., D. J. Weiss, and R. N. Aslin. 2008. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science* 11:122–134.

Mielke, J., A. Baker, and D. Archangeli. 2010. Variability and homogeneity in the english /ɹ/ allophony and /s/ retraction. In C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée, eds., *Laboratory Phonology 10*, pages 699–730. Mouton de Gruyter.

Mielke, J., K. Nielsen, and L. V. Magloughlin. 2013. Phonetic imitation by individuals with Autism Spectrum Disorders: Investigating the role of procedural and declarative memory. *Proceedings of Meetings on Acoustics* 19:060142.

Miller, G. 1983. Informavores. In F. Machlup and U. Mansfield, eds., *The study of information: Interdisciplinary messages*, pages 111–113. Wiley-Interscience.

Minsky, M. L. and S. A. Papert. 1969. *Perceptrons: An introduction to computational geometry*. MIT Press.

Mitchell, T. M. 1980. The need for biases in learning generalizations. Tech. Rep. Technical Report CBM-TR-117, Rutgers University, Rutgers, NJ.

Mitroff, S. R., D. I. Simons, and D. T. Levin. 2004. Nothing compares 2 views: Change blindness can occur despite preserved access to the changed information. *Perception and Psychophysics* 66:1268–1281.

Moreton, E. 2008a. Analytic bias and phonological typology. *Phonology* 25:83–127.

Moreton, E. 2008b. Modelling modularity bias in phonological pattern learning. *West Coast Conference on Formal Linguistics* 27:1–16.

Mowrey, R. and W. Pagliuca. 1995. The reductive character of articulatory evolution. *Rivista di Linguistica* 7:37–124.

Nesset, T. 2008. *Abstract phonology in a concrete model. Cognitive Linguistics and the morphology-phonology interface*. Mouton de Gruyter.

Niyogi, P. 2006. *The computational nature of language learning and evolution*. Cambridge, MA.

Norris, D. and J. M. McQueen. 2008. Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115:357–395.

Nosofsky, R. M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115:39–57.

Nosofsky, R. M. 1988. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory and Cognition* 14:54–65.

Ohala, J. J. 1978. Southern Bantu vs. the world: The case of palatalization of labials. *Berkeley Linguistics Society* 4:370–386.

Ostrand, R., S. E. Blumstein, and J. L. Morgan. 2011. When hearing lips and seeing voices becomes perceiving speech: Auditory-visual integration in lexical access. *Proceedings of the Annual Meeting of the Cognitive Science Society* 33:1376–1381.

Pauen, S. 2002. The global-to-basic shift in infants' categorical thinking: First evidence from a longitudinal study. *International Journal of Behavioral Development* 26:492–499.

Perfors, A. 2011. Memory limitations alone do not lead to over-regularization: An experimental and computational investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society* 33:3274–3279.

Phillips, B. S. 2001. Lexical diffusion, lexical frequency, and lexical analysis. In J. Bybee and P. Hopper, eds., *Frequency and the emergence of linguistic structure*, pages 123–136. John Benjamins.

Piantadosi, S. T. 2014. Zipfs word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin and Review* OnlineFirst.

Pierrehumbert, J. 2001. Why phonological constraints are so coarse-grained? *Language and Cognitive Processes* 16:691–698.

Pierrehumbert, J. 2006. The statistical basis of an unnatural alternation. In L. Goldstein, D. H. Whalen, and C. Best, eds., *Laboratory phonology 8: Varieties of phonological competence*. Mouton de Gruyter.

Pinker, S. and A. Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193.

Plaisted, K., M. ORiordan, and S. Baron-Cohen. 1998. Enhanced visual search for a conjunctive target in autism: A research note. *Journal of Child Psychology and Psychiatry* 39:777–783.

Pycha, A., P. Nowak, E. Shin, and R. Shosted. 2003. Phonological rule-learning and its implications for a theory of vowel harmony. *West Coast Conference on Formal Linguistics* 22:423–435.

Reiss, C. 2004. Constraining the learning path without constraints, or the OCP and NoBanana. In B. Vaux and A. Nevins, eds., *Rules, constraints, and phonological phenomena*. Oxford University Press.

Revill, K. Pirog, R. N. Aslin, M. K. Tanenhaus, and D. Bavelier. 2008. Neural correlates of partial lexical activation. *Proceedings of the National Academy of Sciences* 105:13111–13115.

Richtsmeier, P. 2011. Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology* 2:157–183.

Rogers, T. T. and J. L. McClelland. 2004. *Semantic cognition: A Parallel Distributed Processing approach*. MIT Press.

Rowland, C. F., F. Chang, B. Ambridge, J. M. Pine, and E. V. M. Lieven. 2012. The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition* 125:49–63.

Sankoff, D. and W. Labov. 1979. On the uses of variable rules. *Language in Society* 8:189–222.

Sarle, W. S. 1994. Neural networks and statistical models. *Proceedings of the Annual SAS Users Group International Conference* 19:1–13.

Schane, S., B. Tranel, and H. Lane. 1975. On the psychological reality of a natural rule of syllable structure. *Cognition* 3/4:351–358.

Simon, H. A. 1955. On a class of skew distribution functions. *Biometrika* 42:425–440.

Skousen, R. 1989. *Analogical modeling of language*. Kluwer.

Smoke, K. L. 1932. *An objective study of concept formation*, vol. 42 of *Psychological Monographs*.

Smolensky, P. 1999. Grammar-based connectionist approaches to language. *Cognitive Science* 23:589–613.

Spencer, J. P., S. Perone, L. B. Smith, and L. K. Samuelsson. 2011. Learning words in space and time: Probing the mechanisms behind the 'suspicious coincidence'. *Psychological Science* 22:1049–1057.

Stanners, R., J. Neiser, W. Hernon, and R. Hall. 1979. Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior* 18:399–412.

Stave, M., A. Smolek, and V. Kapatsinski. 2013. Inductive bias against stem changes as perseveration: Experimental evidence for an articulatory approach to output-output faithfulness. *Proceedings of the Annual Meeting of the Cognitive Science Society* 35:3454–3459.

Strobl, C., A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9:307.

Stump, G. T. 2001. *Inflectional morphology*. Cambridge University Press.

Swingley, D. 2007. Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Science* 43:454–464.

Tagliamonte, S. A. and R. H. Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24:135–178.

Thompson, R. F. 2009. Habituation: A history. *Neurobiology of Learning and Memory* 92:127–134.

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

van den Bosch, A. and W. Daelemans. 2013. Implicit schemata and categories in memory-based language processing. *Language and Speech* 56:309–328.

Wedel, A., A. Kaplan, and S. Johnson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 126:179–186.

Weinreich, U., W. Labov, and M. I. Herzog. 1968. *Empirical foundations for a theory of language change*. University of Texas Press.

Westermann, G. and N. Ruh. 2012. A neuroconstructivist model of past tense development and processing. *Psychological Review* 119:649–667.

White, J. 2014. Evidence for a learning bias against saltatory phonological alternations. *Cognition* 130:96–115.

Wierzbicka, A. 1990. Prototypes save. In S. L. Tzohatzidis, ed., *Meanings and prototypes: Studies in linguistic categorization*. Routledge.

Wilson, C. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:945–982.

Wright, R. 2004. A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, and D. Steriade, eds., *Phonetically-based phonology*, pages 34–57. Cambridge University Press.

Xu, F. and J. B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review* 114:245–272.

Yang, C. D. 2002. *Knowledge and learning in natural language*. Oxford University Press.

Yu, A. C. L. 2010. Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PLoS One* 5:e11950.

Yule, G. U. 1944. *The statistical study of literary vocabulary*. Cambridge University Press.

Zuraw, K. 2000. *Patterned exceptions in phonology*. Phd dissertation, UCLA.