Understanding the roles of type and token frequency in usage-based linguistics

Vsevolod Kapatsinski
University of Oregon

Forthcoming in M. Diaz-Campos & S. Balasch (Eds.), *The handbook of usage-based linguistics*. Wiley.

Abstract: This chapter contextualizes token and type frequency effects on accessibility, chunking, productivity and pronunciation in the context of probabilistic, connectionist and analogical views of learning. A review of the effects of frequency in language is provided. It is argued that token frequency as a measure of experience with a particular unit and type frequency as a measure of contextual variability are indispensable for understanding the effects of experience on language structure and linguistic behavior.

Keywords: token frequency, type frequency, productivity, reduction, lexical access, language production, language change, chunking, parallel processing, usage-based

1. Introduction and Chapter Summary

The fundamental premise of usage-based linguistics is that language structure emerges from language use. Consequently, quantifying a person's experience with language is of central importance. Type and token frequency measures are ways to quantify linguistic experience, which rely on the notion that such experience can be discretized into events or units, including words, morphs and segments at the form level, and (less commonly) concepts or semantic features at the level of meaning. Most research has examined the effects of form frequencies, in part because – compared to meanings – forms are more easily operationalized and reliably detected by an analyst. Token frequency then refers simply to the number of times a unit was experienced, and is often operationalized as the number of times it is observed in some corpus (a collection of speech or text). A particular experienced instance of a unit is referred to as a *token* of that unit. For example, there are three tokens of the indefinite article *a* in the preceding sentence, which means that the token frequency of the word *a* in that sentence is 3.

Type frequency relies on the additional assumption that some tokens are perceived as being the same *type*. Type frequency is the number of distinct types that exemplify a certain pattern. The primary use of type frequency is to predict the likelihood of the pattern being extended to new types, i.e., the pattern's *productivity* (Bybee 1985, 1995, 2001). For example, the sentence *"a particular experienced instance of a unit is referred to as a token of that unit*" contains eight tokens of the letter <*a*>, hence its token frequency is 8. However, if we treat words as types, then the type frequency of the letter <*a*> is only 5 because it occurs in 5 distinct words (*a*, *particular*, *instance*, *as*, and *that*). We can use this estimate of type frequency to predict the productivity of <*a*>, i.e., the likelihood that a new English word would be to contain the letter <*a*>. We do this by dividing the type frequency of <*a*> by the total number of types observed, i.e., the size of the lexicon. The sentence above contains the distinct words *a*, *particular*, *experienced*, *instance*, *of*, *unit*, *is*, *referred*, *to*, *as*, *token*, and *that*, for a lexicon size of 12. Given this experience, the best estimate of how likely a new word from the same population would have <a> is 5/12 = 42% because 5 out of the 12 distinct experienced words contain <*a*>. Of course, this assumes that all

experienced words are equally relevant for estimating the behavior of the new word. In reality, words of similar phonology, meaning, and token frequency are actually more relevant than less similar words. Therefore, type frequency must be combined with measures of similarity to predict the likelihood of generalizing a pattern to a new word (e.g., Ernestus and Baayen 2003; Hayes et al. 2009; Olejarczuk and Kapatsinski 2018; Wang and Derwing 1994).

Defining type frequency requires identifying a pattern whose type frequency we are calculating (here, the letter <*a*>), and defining what constitutes a type. Types are usually defined at the level of generality at which we want to predict generalization behavior. Usually, this is the level of words: we want to predict generalization to novel words, and therefore define types as distinct words. For this reason, type frequency is often informally called *dictionary frequency* – the number of distinct entries that would exemplify the pattern in a large dictionary (Bybee 2001). However, types can also be defined at any other level. For example, we could define types as immediate letter contexts around <*a*>. In that case, the type frequency of <*a*> in the same sentence is 7 (the distinct contexts are <_ >, <*p*_*r*>, <*l*_*r*>, <*t*_*n*>, <_ >, <_*s*>, <*h*_*t*>). This definition of *type* is sensible if we want to use the type of frequency of <*a*> to predict generalization to new letter contexts, i.e., if we are interested in the following research question: "if we were to encounter a new sequence of three letters, what is the probability that the middle letter is <*a*>?" Types could also be defined at the level of utterances or sentences, or even entire texts. Type frequency at this level is usually called *contextual diversity* in language processing (following Adelman et al. 2006) or *range* in corpus linguistics (Gries 2013). However, words are in a kind of sweet spot in the hierarchy of linguistic units. On the one hand, words are large enough for speakers to often encounter new ones. As a result, speakers often need to generalize to a new word, compared to a new segment or letter context. On the other hand, words are small enough to recur, rendering type and token frequency more distinct than they are for larger types.

When distinct words are considered types, there is often uncertainty regarding whether morphologically related words should be treated as the same type. For example, do *stand~stood* and *understand~understood* both contribute to the type frequency of an $æn_{PRES}$~$ʊ_{PAST}$ schema in English? In other words, are *stand~stood* and *understand~understood* two *distinct* instances of the $æn_{PRES}$~$ʊ_{PAST}$ schema? And what do we mean by *distinct*? Clearly, the two verbs are not the same – they are both perceptually and semantically distinct. However, they are usually considered the same type for the purposes of predicting the productivity of the past tense pattern they exemplify (Albright and Hayes 2003; Bybee 1995) because *understand* does not provide independent evidence for the productivity of $æn_{PRES}$~$ʊ_{PAST}$. The (usually implicit) argument is as follows. Prefixed verbs always behave like their base verbs with respect to the choice of the past tense schema in English (see Bybee 1995, for the same argument for German). Therefore, knowing that the past tense of *understand* is *understood* provides no additional information about the likelihood of a novel verb exemplifying the same pattern, as long as the learner already knows that the past tense of *stand* is *stood*.

From this *inferential* perspective, tokens constitute distinct types to the extent that they are *independently informative* about the behavior of novel types. That is, we can say that a token (*t*) is a new type (*T*), for the purposes of estimating the productivity of some schema (*S*), when the learner would not know (for sure) whether *t* exemplifies *S* before encountering it. In other words, the conditional probability $p(S|t)$ should change after encountering *t*.

Notice that the grouping of tokens into types is not context-independent: a type is only a type for the purposes of some prediction task. Thus, even though *understand* and *stand* might be the same type for predicting past tense, they constitute distinct types for predicting the productivity of the stem *stand* or recognizing it in a new context (type frequency of a stem is often called *morphological family size* following Schreuder and Baayen 1997).

In what follows, I describe the effects of type and token frequency reported in the literature (Section 2), the current approaches to modeling these effects (Section 3), and highlight some of the pending issues (Section 4). As we will see, even though both type and token frequency have reliable behavioral correlates, these effects can be modeled in several very distinct ways. Progress will likely come from examining how the effects of type and token frequency vary across distinct tasks and levels of analysis, and determining whether type frequency, predictability, or variability underlies apparent type frequency effects.

2. Background on the relevant literature in the area

2.1. The effects of token frequency in psycholinguistics

Accessibility of a unit can be defined as the ease with which it can be perceived or produced. Most token frequency effects can be thought of as effects on accessibility. For example, high token frequency of a word makes it easier to recognize (Howes 1957; Howes and Solomon 1951) and produce (Oldfield and Wingfield 1965). This effect of token frequency has been most commonly modeled as a difference in resting activation levels of lexical representations residing in a parallel processing system (following Morton's 1969 *logogen model*). Perceptual input activates all words that partially match it; the words then compete for recognition, and the word that is activated most strongly wins. Frequent words have stronger activation levels in the absence of perceptual input, and therefore require less input activation (a lower degree of similarity to the signal) to win the competition.

An important constraint on models of word recognition is the *neighborhood frequency effect* on reaction times (Grainger et al. 1989; Luce and Pisoni 1998): words that have *neighbors*, or phonologically-similar competitors, of high token frequency are recognized more slowly than words whose competitors are rare. For example, a cat lover might have a harder time recognizing the word *cad*, because of the high frequency of the word *cat* in their experience. In the logogen model (Morton 1969), words were racing to reach an activation threshold independently. Therefore, the time needed for a word to reach the threshold was predicted to be independent of how strong its competitors are. The neighborhood frequency effect on reaction times can be accounted for by several processing mechanisms, including lateral inhibition between competitors (the more activated a word is, the more it is able to inhibit competing words; McClelland and Rumelhart 1981), and division of activation coming from the signal between the matching words (Kapatsinski 2006). These effects can also be accounted for by competition at the timescale of learning rather than processing. In particular, all models of recognition associate sublexical cues to lexical or semantic representations. Competition can be modeled by assuming that when [k] cues *cat*, its association with *cat* grows, but, crucially, its associations with all other word weaken (Allopenna et al. 1998; Baayen 2011; Baayen et al. 2016). At a deeper level,

these alternative mechanisms implement the idea that words must compete for probability mass: because only one word must be present in the signal at a certain time, the probabilities of all words given a signal must sum to one, hence evidence for one word is evidence against other words being present at the same time (Norris and McQueen 2008).

2.2. The effects of accessibility on linguistic structure

The influence of token frequency on accessibility of a form has many consequences for language change, and the emergence of linguistic structure. First, high token frequency makes a form resistant to analogical change, i.e., change that is due to analogy to other forms (Bybee 1985, 2001; Bybee and Brewer 1980; Carroll et al. 2012; Hay et al. 2015; Lieberman et al. 2007; Phillips 1984, 2001). That is, infrequent words are more likely to fall in line with the rest of the lexicon. For example, Lieberman et al. (2007) show, for English, that "a verb that is 100 times less frequent regularizes 10 times as fast" (p.713). Hay et al. (2015) show that changes in the pronunciation of front lax vowels in New Zealand English over the last 100 years were led by low-frequency words, a result that suggests an analogical mechanism for the advancement of this chain shift (Bybee 2001; Phillips 2001; though cf. Hay et al. for an alternative).

Zipf (1949) proposed that token frequency of a form can lead that form to be extended to additional uses. He showed that the token frequency of a word correlates with the number of senses it has in a dictionary. Harmon and Kapatsinski (2017) proposed and tested a mechanistic account of this effect by pointing out that token frequency should increase accessibility of a form not only in the contexts in which it has been encountered but also in similar contexts. In production, a form is activated by a distributed semantic representation that shares features with the meanings of other forms. This leads forms to compete for production even when they are not exactly synonymous (see also Burke et al. 2004; De Smet 2016; Ferreira and Griffin 2003; Kapatsinski 2009; Koranda et al. 2021; Srinivasan and Winter 2021). The higher the token frequency of a form, the stronger its association with its semantic features. This makes the form more accessible when the speaker intends to express the original meaning of the form, leading it to be resistant to analogical change. It also makes it more accessible when the speaker intends to express related meanings, i.e., meanings that share only some features with the meaning(s) with which the form was previously experienced. When a speaker produces a form to express a meaning which it has not been used to express before, she plants a seed for a semantic extension. If accepted by the speech community, it grows into a semantic change. The outcome of this process is that frequent forms are used in a wide range of uses that often have no features in common but form a chain of family resemblances (e.g., Lakoff 1987; Plaster and Polinsky 2010). Harmon and Kapatsinski (2017) show that the effect of frequency on semantic extension is mediated by accessibility: when frequent and infrequent forms are made equally accessible, there is no preference to extend frequent forms. These results show that the likelihood of using a form to express a particular meaning depends both on the similarity between that meaning and the known meaning(s) of the form, and on the accessibility of the form, which is influenced by token frequency.

A special case of this effect is represented by the leveling of morphological paradigms in favor of frequent forms (Bybee and Brewer 1980; Tiersma 1982). Morphological paradigms are sets of word forms that show extreme semantic similarity. For example, *cat* and *cats* share almost every

aspect of their meaning, except for plurality. It is therefore to be expected that one of the forms might be extended to replace the other, or might be activated enough during the production of the other form to be *blended* with it, changing its shape (Kapatsinski 2013). Tiersma (1982) showed that frequent forms indeed reshape or replace less frequent forms in paradigm leveling. Across Frisian, Slavic, Hebrew, and German, most plural nouns were reshaped by singular forms, except for those whose referents usually come in groups, such as geese. In those nouns, the singular was reshaped by the plural. Bybee and Brewer (1980) showed that semantic similarity between the forms also matters: forms whose semantics are more similar are more likely to influence each other than forms whose semantics are dissimilar. Both effects are expected from distributed semantic representations activating associated forms in parallel, with frequent forms having strongest associations or higher resting activation levels (see also Hoeffner and McClelland 1993, for analogous effects in child language).

An additional mechanism by which high token frequency can cause semantic change is represented by bleaching through *habituation*, i.e., the process by which a repeated stimulus loses its ability to evoke the associated response (Bybee 2003). Habituation is likely behind the frequent rise and fall of intensifiers: an overused intensifier loses its oomph, requiring reinforcement by an additional intensifier or replacement by an alternative. It can also explain the need to reinforce diminutives with additional diminutives in languages like Russian. Habituation by itself appears unlikely to explain the changes that result in polysemous frequent words, both because such words do not have one vague meaning but rather a network of specific uses, and because frequent forms do not appear to lose their ability to evoke *all* features of an associated meaning. Indeed, in judgment tasks, experience with a word generally makes listeners restrict the word to the meanings or contexts with which it has been experienced (Harmon and Kapatsinski 2017; Xu and Tenenbaum 2007a; Theakston 2004). That is, repetition generally does not make the word dissociate from its meaning, except for any parts of the meaning that constitute an affective response to an unexpected stimulus. For example, repeating the word *kitty* does not make it less effective at referring to cats but can make it less effective at eliciting the associated feelings (Harmon and Kapatsinski 2017; cf., Bybee 2003).

2.3. Type frequency and accessibility

Type frequency has also been observed to affect accessibility, under other names. First, words with a large *morphological family size* are easier to recognize (Bertram et al. 2000; De Jong et al. 2000; Moscoso del Prado Martín et al. 2004b; Schreuder and Baayen 1997). Morphological family size refers to the number of distinct words that share the stem with the word in question. This effect can therefore be understood as type frequency of the stem (De Jong et al. 2000).

Second, Adelman et al. (2006) have argued that word recognition is predicted by a kind of word type frequency, the number of distinct documents or contexts in which the word occurs, which they called *contextual diversity* (see Gries, Chapter 33, this volume, for discussion). The superiority of contextual diversity measures over token frequency measures derived from natural corpora has been disputed, and the question is difficult to address definitively because token frequency and document count are strongly correlated ($r > .95$; Brysbaert and New 2009). However, Jones et al. (2012) have manipulated frequency and diversity independently in an experiment, and showed that the diversity of contexts in which a word occurs can independently

influence word recognition. They also showed that it is not just the count of distinct contexts that matters but the variability of the contexts: if a word occurs in very similar documents, it is less accessible than a word that occurs in a wide variety of documents. Adelman and colleagues related contextual diversity effects to Anderson and Schooler's (1991) Principle of Likely Need, which proposes that a word that occurs in a wide variety of contexts is more likely to occur in a new context.

2.4. Token and type frequency effects on chunking

High token frequency of a structure can result in the parts of that structure fusing together into a chunk (see also Beckner, Chapter 18, this volume). As Bybee (2002b) writes, "units used together fuse together". For example, English auxiliaries fuse together with frequent subjects; particularly, pronouns. This happens despite the fact that, semantically, the auxiliary belongs with the following verb rather than the preceding noun phrase (Bybee 2002b).

Many effects of token frequency on chunking can be understood as either token frequency of a unit strengthening a holistic representation of that unit, or as strengthening associations between smaller units (Baayen et al. 2013; Taft 2004). For example, the token frequency of a morphologically complex word like *cats* influences its recognition and production above and beyond the frequency of its stem (Alegre and Gordon 1999; Baayen et al. 2003; Bybee 2002a; Kapatsinski 2010b), a finding that also extends to larger compositional units (Arnon and Snider 2010; Tremblay and Baayen 2010). However, the rapid recognition or production of *cats* could be accounted for by how easily the context *cat* activates *-s* (but see the discussion of top-down masking below).

However, evidence for token frequency strengthening a holistic perceptual unit comes from the effect of top-down masking: a form is harder to detect when it occurs inside a frequent larger unit (Giroux and Rey 2009; Healy 1976, 1994; Kapatsinski 2021b; Kapatsinski and Radicke 2009; Vogel Sosa and MacFarlane 2002). For example, Kapatsinski and Radicke (2009) find that *up* is harder to detect inside of the most frequent verb-particle combinations like *come up*. This effect suggests that psychologically real units can compete with their parts for activation, and that a frequent unit is likely to grow *autonomous* and non-compositional, recognized and produced independently of its parts (Blumenthal-Dramé et al., 2017; Bybee 1985, 1995; Bybee and Brewer 1980; Hay 2001; Moder 1992). Additional evidence for this hypothesis is provided by Moder (1992), who showed that frequent verbs do not prime the past tense patterns they exemplify as effectively as verbs of medium frequency (see also Blumenthal-Dramé et al., 2017). A possible processing mechanism for top-down masking is that the comprehender "moves on" upon assigning an interpretation to a particular stretch of speech or writing. When a large unit like *come up* is recognized, there is no need to continue processing the corresponding part of the speech stream. Therefore attention is withdrawn, and activation from the signal is shut off, preventing activation of the less accessible component units from rising to the level of conscious awareness. The withdrawal of attention upon recognition is particularly obvious in reading, where it can be observed in eye movements (Greenberg et al. 2004), but may also hold for speech processing.

In production, frequent units appear to be relatively unlikely to be interrupted, either when they need to be replaced (Kapatsinski 2010a), or when the speaker searches for a spot to restart production from (Harmon and Kapatsinski 2021). Harmon and Kapatsinski (2021) explain this effect by proposing that selection of the next unit to produce is driven both by top-down semantic input and by the preceding context, which compete for wiring with the upcoming unit (see also Arnon and Ramscar 2012). It is because of this competition that words occurring in specific contexts are difficult to access outside of the experienced contexts (Jones et al. 2012). The influence of context facilitates production of conventional expressions but makes it more difficult to follow the path less trodden. In rare cases, it means that highly likely words can be produced in error. For example, the former Israeli prime minister Benjamin Netanyahu has recently referred to the UK prime minister Boris Johnson as Boris Yeltzin, the former president of Russia, presumably because *Boris* is an excellent cue to *Yeltzin* in Netanyahu's experience. McCauley et al. (2021) systematically document the existence of such *habit slip* errors, in which a frequent unit replaces a rarer one (see also Beckner 2018).

Unitization of a sequence *ab* is likely to be reduced when one of its elements occurs in other contexts (see also Beckner, Chapter 18, this volume; Gries, Chapter 33, this volume). That is, if *a* or *b* has high type frequency, or high token frequency in other contexts, *ab* should be less of a unit (Gries 2012). For example, it is likely that the strength of *Boris* as a cue to *Yeltzin* increases both because *Boris* is frequently followed by *Yeltzin* and because *Yeltzin* rarely occurs in other contexts. However, although all learning models agree on the importance of the token frequency of *ab*, they disagree on the importance of the frequency with which *b* occurs without *a*, *a* occurs without *b*, and the frequency with which *a* and *b* are absent together (Kapatsinski and Harmon 2017). For example, because *Johnson* is a much more common last name than *Yeltzin*, it often occurs without *Boris*. Do these occurrences weaken the ability of *Boris* to cue *Johnson*? Do they weaken them to the same extent as the occurrences of *Boris Johnson* strengthen the association? Do the occurrences of bigrams that contain neither *Boris* nor *Johnson* strengthen the association between the two words because both are simultaneously absent? These questions are an active area of current research on collocations (e.g., Gries 2013; Schneider 2020). One question that is somewhat overlooked in this domain is whether type frequency matters for unitization, e.g., whether a word that occurs in a greater variety of contexts becomes more autonomous, independently of its token frequency in the current context (Harmon and Kapatsinski 2021).

2.5. Token frequency effects on articulation

Although high token frequency makes a word resistant to analogical change, it also makes the word more susceptible to articulatory streamlining (see also Brown, Chapter 7, this volume; Gradoville, Chapter 31, this volume). This can be seen most clearly in cases of *special reduction*, which refers to streamlining processes that are evident only in a small number of words, inevitably ones of high token frequency, e.g., *probably* being reduced to *prolly* or *I don't know* being reduced to a nasalized schwa and a prosodic contour (Bybee and Scheibman 1999). Importantly, reduction can be conditioned by meaning, suggesting that the units whose production is being streamlined in language production are meaningful, schematic ones. Thus, Bybee and Scheibman (1999) point out that *don't* in *I don't know* can only reduce to a nasal schwa when it functions as an expression of uncertainty, rather than as a literal proposition. Gahl (2008) and Lohmann (2018) show that frequent members of homophone pairs such as *time* and

thyme are pronounced more quickly than their less frequent counterparts. In addition to shortening duration, high token frequency leads to articulatory reduction, understood as the decrease in magnitude of articulatory movement (Mowrey and Pagliuca 1995) and a smoothing of the velocity profile such that deceleration and acceleration are minimized (Kapatsinski 2018a; Sosnik et al. 2004). Reductive changes include vowel centralization and raising in unstressed syllables, assimilation, lenition of closures between vowels, and many others, accounting for a clear majority of sound changes in languages of the world (Bybee and Easterday 2019; Mowrey and Pagliuca 1995). It has been argued that the reductive motivation behind a change can be inferred from finding that it started in frequent words, because these are the words with which the speaker has the most practice (Bybee 2001; Mowrey and Pagliuca 1995).

Based on this work, usage-based linguists have suggested treating words (and other memorized meaningful units) as being subject to *automatization* (Bybee 2002a; Kapatsinski 2010a, 2018a; Kapatsinski et al. 2020; Tomaschek et al. 2018). Automatization is the process by which production practice leads the speaker to be able to produce the same sequence of actions more quickly and robustly, i.e., with less unintended variability (Bryan and Harter 1899). In line with this idea, Tomaschek et al. (2018) found that speakers produce frequent words more quickly without spectral reduction, i.e., without sacrificing the acoustic distinctness of the words' stressed vowels. However, diachronically, temporal reduction precedes and predicts reduction in articulatory movement magnitude (see Cohen Priva 2020, for consonant lenition). It may therefore be the case that spectral reduction will eventually follow temporal reduction. The likely reason for this reduction is that automatization of language production is guided by social feedback, reducing parts of the action sequence that are not essential for accomplishing conversational goals and preserving or enhancing those that are essential (Kapatsinski 2018a; Kapatsinski et al. 2020). Support for this idea comes from experimental studies showing that, when speakers are misunderstood, they enhance acoustic cues that the listener has misperceived (Buz et al. 2016).

This type of *reinforcement learning* process may explain how low-information parts of the speech signal, such as unstressed vowels, become more reduced over time, while high-information parts of the signal such as stressed vowels may become more prominent (Bybee et al., 1998; see also Cohen Priva 2017; Seyfarth 2014; Soskuthy and Hay 2017; Wedel et al. 2013, for information content of words and segments inversely correlating with reduction). Type frequency of a unit can be seen as a simple measure of its information content because, controlling for token frequency, a schema of high type frequency occurs relatively unpredictably. Specifically, predictability of a word in a context is p(word|context), which is the word's token frequency in that context, divided by the frequency of the context (which is independent of the identity of the word). A word's expected probability in a context is the average of its probabilities across contexts, i.e., $\frac{\sum_{context} p(word|context)}{N}$, where N is the number of contexts in which the word occurs. Because the number of contexts is the word's type frequency, the expected probability of a word in context is inversely proportional to its type frequency (controlling for token frequency).

2.6. Type and token frequency effects on productivity

Type frequency is a relatively uncontroversial correlate of *productivity*, i.e., the likelihood of using a form or pattern in a new context (Bybee 1985, 1995). Other things being equal, the likelihood of using a pattern in a new context is proportional to its type frequency in similar known contexts (e.g., Ernestus and Baayen 2003; Hayes et al. 2009; Olejarczuk and Kapatsinski 2018; Wang and Derwing 1994). Much of the research on productivity can be understood as identifying what *similar* mean in the sentence above (e.g., Albright and Hayes 2003). In generative models, the similar contexts form a *classical category*, one defined by a set of necessary and sufficient conditions (e.g., the set of English verbs that end in a voiceless fricative). Because all words that belong to such a category are equal members, the productivity of a pattern they exemplify is exactly proportional to its type frequency (and inversely proportional to the number of types that constitute exceptions; Albright and Hayes 2003). In connectionist and analogical models, the categories have gradient structure, so type frequency interacts with similarity (e.g., Hare et al. 1995; Suttle and Goldberg 2011). For example, in analogical models of morphology, words vote for patterns they exemplify (Daelemans and van den Bosch 2005). Type frequency matters because the greater the number of words voting for a pattern, the more likely it will be to apply to the novel word. However, words that are more similar to the novel word have more votes. Therefore, the influence of each type is weighted by its similarity to the novel word. The end result is that the likelihood of applying a pattern is proportional to type frequency times the average similarity between the types exemplifying a pattern and the novel type.

Proponents of the Dual Mechanism Model in morphology (Clahsen 1999; Clahsen et al. 1992; Marcus et al. 1995; Pinker 1999) argued that type frequency can be dissociated from productivity and therefore does not play a role in it. The primary evidence for this claim is provided by *minority defaults*, patterns that are productive in default contexts despite having a low type frequency. The classic example is the German plural *-s*. However, minority defaults can only provide evidence against an influence of type frequency if type frequency in relevant contexts cannot account for the data. For example, it has been argued that the German *-s* pattern serves as the default only for words ending in full vowels (Köpcke 1988, 1998). If so, it is type frequency amongst such words that is most relevant, and type frequency of *-s* in this subset of the lexicon is relatively high (Bybee 1995; Köpcke 1988, 1998; Yang 2016). Furthermore, approaches that consider type frequency to interact with similarity are not threatened by minority defaults if the types exemplifying the default pattern are more diverse than those exemplifying competitor patterns (Boudelaa and Gaskell 2002; Hare et al. 1995).

There is much disagreement on the role of token frequency in breadth of generalization. Bayesian researchers have argued against such an effect (Perfors et al. 2014; Xu and Tenenbaum 2007a). This proposal is backed up by null results (e.g., Albright and Hayes 2003; Perfors et al. 2014) as well as by finding that the same change in token frequency has an effect only if the tokens are distinct types (Xu and Tenenbaum 2007a). However, token frequency should not matter from a Bayesian perspective only if the average frequency of a type provides zero information about the existence of more types that exemplify the pattern (cf., Baayen 1993). This assumption is only true under some sampling assumptions. For example, suppose that the learner assumes that experienced tokens of a pattern are independently sampled from the population of tokens exemplifying the pattern, and observes that the sampled tokens keep exemplifying the same types. In this situation, the high token frequency of the experienced types provides

evidence that there are no more types that exemplify this pattern. Under this sampling assumption, token frequency should matter for productivity, detracting from productivity of a pattern. Since we know that people are sensitive to sampling assumptions (Xu and Tenenbaum 2007b), demonstrations that token frequency played no role should probably be interpreted as participants adopting a particular sampling assumption that may not hold in real language learning (Kapatsinski 2018b).

Exemplar models, which are analogical models that allow individual tokens to vote for patterns or categories they exemplify, suggest that high token frequency should help a pattern because more tokens vote for it (Nosofsky 1988; see the discussion in Albright and Hayes 2003). Patterns are also usually helped by token frequency in connectionist models, weights exemplifying a token-frequent pattern benefitting from the increased experience (e.g., Moscoso del Prado Martin et al. 2004a). However, empirically, it is fair to say that allowing exemplars to vote is usually unhelpful to predict productivity, because the conserving effect of frequency ensures that high-frequency words are the ones most likely to be exceptional (e.g., Lieberman et al., 2007).

Several researchers have argued that high token frequency detracts from productivity of a pattern. Bybee (1985, 1995, 2001; Bybee and Brewer 1980) suggested that types of high token frequency might be recognized and produced directly, without use of the pattern in question, reducing its productivity, and may not even be associated with the pattern in memory (see also Moder 1992). Baayen (1993) showed that productivity of a pattern is well predicted by the proportion of hapax legomena, words observed only once in the corpus (see also Zeldes 2012, for syntactic patterns). Hapax legomena provide evidence for the pattern being useful to recognize and produce words (or other constructs), because hapaxes are novel and so cannot be produced or recognized directly. Therefore, the proportion of hapax legomena, and low token frequency generally, makes the pattern more likely to be needed in future language use. Therefore, a detrimental effect of high token frequency can be derived from the Principle of Likely Need. Kapatsinski (2021a) argues that language users may implicitly know that frequent words are not exactly like rare or novel words, and therefore would generalize patterns to novel words from other rare words, also resulting in a correlation between low token frequency and productivity. For example, because Latinate words tend to be lower frequency than Germanic words, and are over-represented in academic vocabulary, the typical college student participant in a psycholinguistic experiment is much more likely to encounter a novel Latinate word than a novel Germanic word. They therefore might expect a novel word to bear a Latinate suffix like *-ity*, rather than a Germanic suffix like *-ness*, producing *-ity* more than would be expected from its type frequency. Overall, there is a wealth of correlational data showing that low token frequency correlates with high productivity, but experimental data establishing a causal link is lacking.

Some authors have proposed that token frequency interacts with type frequency or the distribution of tokens over types. Barðdal (2008) proposed that patterns can be extended *either* by analogy or by extension of an abstract schema, and therefore high token frequency helps patterns of low type frequency but hurts those of high type frequency. Goldberg et al. (2004) emphasized the need to associate an abstract pattern with a meaning, and reasoned that a skewed, Zipfian distribution of token frequency is particularly helpful for learning productive syntactic constructions: the high-frequency types allow the language learner to acquire the meaning of the pattern, whereas low-frequency types allow it to maintain productivity. Madlener (2016) has

argued that the Zipfian distribution is helpful only if type frequency is high enough because enough low-frequency types need to be learned for the pattern to be productive. However, the evidence for an influence of token frequency distributions within patterns is rather inconclusive (Madlener 2016).

3. Current Approaches

Current approaches to the effects of experience on language can be classified as falling into three traditions: probabilistic, analogical, and connectionist models. The probabilistic approach makes use of structured, interpretable, often hierarchical representations that are assigned explicit probabilities through experience. The currency of the mind in this approach is probability rather than activation or similarity, thus token and type frequencies of linguistic units play a direct role in learning. The goal of learning is to infer a model of how the experienced utterances were generated, usually through the use of Bayesian inference (e.g., Griffiths et al. 2010; Kapatsinski 2021a; O'Donnell 2015; Perfors et al. 2014; Xu and Tenenbaum 2007a, 2007b; but see also Albright and Hayes 2003; McCauley and Christiansen 2019, for examples of non-Bayesian probabilistic models operating on interpretable units). A significant strength of this approach is that it allows for explicit representation of linguistic hierarchies, which allows the learner to form hypotheses and beliefs at multiple levels of generalization (e.g., Kapatsinski 2021a; see also Divjak & Milin, Chapter 19, this volume). An important disadvantage is that the space of possible representations is completely unconstrained, which tempts the modeler to build solutions into innate representations rather than letting structure emerge (McClelland et al. 2010; see also Nixon & Tomaschek, Chapter 9, this volume).

The analogical approach eschews abstraction, so that linguistic experience is represented by a cloud of stored instances (see Ambridge 2020; Daelemans and van den Bosch 2005; Divjak & Milin, Chapter 19, this volume; Johns and Jones 2015; Goldinger 1998; Jamieson et al. 2012; Nosofsky 1988; Racz et al. 2020; Skousen 1989). Rather than storing generalizations about what ought to be done in various situations, an analogical learner allows the stored experiences of situations (*exemplars*) to vote for what to do when a particular situation presents itself. The exemplars are situated in a multidimensional similarity space that determines how much any given exemplar influences the choice in a particular situation. By eschewing abstraction, an analogical perspective has traditionally rejected the notion of a type, but analogical models in usage-based linguistics have reintroduced types into this approach (Bybee 2001; Kapatsinski 2006; Pierrehumbert 2001). Because similarity between exemplars fully determines their mutual influence, the currency of the mind is similarity rather than probability or activation. The analogical approach is well suited to modeling the existence of radial categories that defy a simple featural description (e.g., Bybee and Eddington 2006; Medin & Schaffer 1978), and allows the learner to make predictions without learning what the strongest influences on a choice, or most predictive features of situations are. It represents the possibility that learners perceive patterns holistically rather than decomposing them into individualizable features. Learning is also much simpler with this approach, because it involves simply storing experiences rather than trying to infer how important the various features of those experiences are (e.g., Jamieson et al. 2012). However, in the absence of stored generalizations about feature weights, it can face difficulties when different weightings of the dimensions are needed to make different choices (Kapatsinski 2014). It may also be overly susceptible to analogies based on a single type, especially one of high token frequency (Albright and Hayes 2003).

Like the analogical approach, the distributed connectionist approach also eschews storing complex structured representations (e.g., phonemes or words; see Nixon & Tomaschek, Chapter 9, this volume; Divjak & Milin, Chapter 19, this volume). Instead of operating on such interpretable units as elements, connectionist models learn associations between simple neuron-like processing elements that simply convert input activation into output activation by passing it through some function (e.g., Baayen et al. 2011; Hare et al. 1995; Harmon and Kapatsinski 2021; Hoppe et al. 2020; Rogers and McClelland 2004; Kapatsinski 2021b). Linguistic representations are distributed patterns of activation and connectivity over elements that do not themselves have a linguistic interpretation. The goals of learning are different from those of a probabilistic learner. First, a connectionist model aims for accurate and timely prediction, whereas a probabilistic model often aims to discover the true causal structure of the world. Second, connectionist models aim only to replicate of adaptive behaviors in the right contexts, not necessarily generating them in the same way they were generated by others. Instead, probabilistic models seek knowledge of how the behaviors are generated to replicate the generation process (e.g., O'Donnell 2015). Connectionist models differ from analogical models in that the ideal representational elements of a connectionist model are 'microfeatures' such as the acoustic energy at a certain frequency at a certain time rather than rich memories of situations: the smallest units rather than the largest units are the most basic (Hoppe et al. 2020). In a connectionist approach, there are no linguistic units, exemplars, schemas or categories explicitly represented (e.g., Rogers and McClelland 2004). There are therefore no explicit representations of types, or individual tokens of experience, although type and token frequency can strongly correlate with the variables that the network is actually tracking (e.g., Harmon and Kapatsinski 2021). The network's behavior can therefore be described in terms of type and token frequency but only approximately. The mind deals in activation and inhibition rather than probability, and similarity emerges from overlap between distributed activation patterns.

The connectionist approach represents the current state of the art in computational linguistics and artificial intelligence, but (as with the analogical approach), there are questions regarding whether it is overly influenced by token frequency compared to type frequency in generalizing to novel word (e.g., Moscoso del Prado Martín et al. 2004a, Simulation 1, show that ignoring token frequency results in much better model performance on novel words), and whether language is more structured and categorical than connectionist networks suggest (Griffiths et al. 2010). In particular, recent work suggests that even the most powerful neural networks can have difficulties in producing humanlike generalization to novel items (Corkery et al. 2019; Liu and Hulden 2021; McCurdy et al. 2020).

4. Pending Issues

Despite the wealth of results reviewed above, many questions about the roles of type and token frequency require empirical attention. First, most research on type frequency has examined its influence on generalization. However, it is quite likely that type frequency also influences how familiar instances of the pattern are processed and pronounced by influencing how independent the pattern is from the context (e.g., Álvarez et al., 2001; Harmon and Kapatsinski 2021). Conversely, the effects of token frequency on generalization are quite inconsistent, and the

interaction between the token frequency distribution and type frequency needs to be explored in greater detail (Madlener 2016).

A central question separating probabilistic and connectionist models is whether type frequency has a *direct* influence on productivity (e.g., Albright and Hayes 2003; Perfors et al. 2014; Xu and Tenenbaum 2007a; Yang 2016), or if it is only an imperfect, approximate measure of variability or predictability (Baayen et al. 2011, 2013, 2016; Hare et al. 1995). It may be possible to induce the learners to perceive the same set of tokens as consisting of a single type or of multiple types by manipulating the instructions to ensure that identical tokens are perceived as either repetitions of the same type, or tokens of different types that happen to look or sound the same (e.g., Hsu and Griffiths 2010; Xu & Tenenbaum, 2007b). If such a manipulation influences the effect of the tokens on generalization, this would go help establish type frequency as an independent influence on behavior distinct from variability, as proposed by probabilistic models.

In lexical processing, there has been disagreement on whether token frequency should be considered an independent influence on processing (e.g., Allopenna et al. 1998; Brothers and Kuperberg 2021; Plaut and Booth 2000), or if it should be taken to be an imperfect measure of something else (e.g., *contextual diversity*, as the number of distinct contexts in which the unit has been experienced, Adelman et al. 2006; or *age of acquisition*, the age at which the word was first learned, Morrison et al., 1992) or word structure (Landauer and Streeter 1973). However, it is now clear that frequency can influence processing above and beyond these additional predictors (e.g., Baayen et al. 2016; Goldiamond and Hawkins 1958; Juhasz et al. 2019). The remaining question is whether frequency should be considered as an imperfect measure of *predictability* in context (Baayen 2010; Baayen et al. 2016; Brothers and Kuperberg 2021; Gries 2012; Jurafsky et al. 2001). It is very tempting to consider frequency as simply to an imperfect measure of predictability (e.g., Jurafsky et al. 2001). However, recent work has suggested that the two effects may be independent, and may have different functional forms (Brothers and Kuperberg 2021; Goodkind and Bicknell 2021). In some models, the two also have distinct loci: connection strength for predictability and resting activation level for token frequency.

In my view, token frequencies should continue to be used as predictors of linguistic behavior even if they turn out to be merely imperfect correlates of predictability (cf., Gries 2012; Chapter 33, this volume). First, predictability either *is* a function of probability, or is well approximated by a linear combination of probabilities (e.g., Allan 1980; Ellis 2006). For example, Harmon and Kapatsinski (2021) show that predictability of a word to a state-of-the-art deep recurrent network, which predicts words from preceding contexts of unlimited length, turns out to largely reduce to its probability given the immediately preceding word ($r > 0.9$). That is, predictability to the network is largely reducible to transitional probability, the probability of a word given its predecessor. Transitional probability, like other probabilities, can be decomposed into a set of frequencies. Thus, the probability of *cat* given a preceding *the*, $p(cat|the\_\_) = freq(the\ cat)/freq(the)$. Because most behavioral measures are better correlated with logarithmically transformed frequencies or probabilities than with their raw counterparts (Figure 1), let us log transform this equation into $log(p(cat|the\_\_)) = log(freq(the\ cat)) - log(freq(the))$. Thus, log transitional probability of *cat* given *the* (also known as *surprisal*) is simply the log frequency of *the cat* minus the log frequency of *the*. That is, surprisal is a linear combination of two log token frequencies. Thus, when modeling some behavior with a linear (regression) model, we have a

choice. We could either use surprisal as a predictor or instead use the two component log token frequencies. The token frequencies will always account for at least as much variance in behavior. By using surprisal, we assume the effects of the two token frequencies to be equal and opposite. By using the component frequencies instead, we can let the data speak to whether this assumption is true, i.e., whether a token of experience with *the cat* matters as much as a token of experience with *the*. Experimental research on contingency learning suggests that this assumption is likely not to hold because encountering a unit (here *cat*) generally shifts beliefs more than not encountering it does (Kao and Wasserman 1993; Kapatsinski and Harmon 2017). In other words, accessibility of *cat* after *the* is likely to be increased after encountering *the cat* more than it would be decreased by encountering *the* followed by some other, non-feline word. If this is true, using conditional probability instead of the two component token frequencies would result in a poorer model of the behavior. More importantly, however, using token frequencies as predictors is often preferrable because it provides an estimate of the importance of each distinct type of experience, in a way that more complex predictors that combine frequencies do not.

The influence of token frequency on pronunciation has been described in terms of automatization. However, the literature on automatization is divided on the relationship between automaticity and flexibility (e.g., Bilalic et al. 2008; Du et al. 2021). In the context of language production, we can ask whether speakers have less or more control over the production of frequent words (Kapatsinski et al. 2020; see also Tantucci and Di Cristofaro 2020). Following Bryan and Harter's (1899) original work on automaticity, Kapatsinski et al. (2020) proposed the working hypothesis that intentional variability – such as the effects of emphasis – should increase with experience with a unit (while unintended variability decreases). However, little empirical research on the question exists. An interesting follow-up question is whether token frequency interacts with contextual and information-structure variability: we might expect the production of frequent words to be more flexible as long as speakers have practice producing the word in multiple ways.

Another pressing issue in this domain is whether increased token frequency always favors reduction. As shown by Bybee (2002a) and Raymond and Brown (2012), reduction is particularly favored by frequency of occurrence in reduction-favoring contexts (see Brown, Chapter 10, this volume). However, it is not clear how this effect interacts with automatization caused by increasing token frequency. Specifically, would additional tokens of occurrence in a reduction-disfavoring context increase reduction (because of automatization) or decrease it (because of accumulation of unreduced exemplars)?

Similarly, the interaction between the conserving effect of token frequency on analogical change and the conducive effect of token frequency on reductive change has not been explored empirically. At least implicitly, previous work has assumed that analogical changes and reductive changes are mutually exclusive, i.e., that reductive changes do not spread by analogy. However, this assumption could well be incorrect, and needs to be empirically tested (Kapatsinski 2021a).

5. Final remarks

Token frequency is a measure of experience, calculating how often a language user has a particular experience. Effects of token frequency are therefore a crucial window on how linguistic representations change as a result of language experience and language use. Type frequency is, in turn, a simple measure of variability, inversely proportional to the average token frequency across types. Token and type frequencies can, and often should, be combined into more complex measures, such as conditional probability and entropy (see also Gries, Chapter 33, this volume; Turnbull, Chapter 8, this volume). It is essential to continue building explicit probabilistic, analogical and connectionist models in which frequency effects may fall out from basic, well-grounded assumptions about learning, processing and representation (e.g., Divjak & Milin, Chapter 17, this volume; Nixon & Tomaschek, Chapter 9, this volume). We should also use simple token and type frequency predictors to understand the behavior of such models, which are often strongly correlated with these simple measures (e.g., Harmon and Kapatsinski 2021). Much empirical work on frequency effects is being done, and much remains to be done. It is to be hoped that the focus on the effects of experience on linguistic behavior and language structure will continue intensifying.

References:

Albright, A., and Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2): 119-161.

Alegre, M., and Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40(1): 41-61.

Allan, L.G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society* 15: 147–149.

Allopenna, P.D., Magnuson, J.S., and Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38(4): 419-439.

Álvarez, C. J., Carreiras, M., & Taft, M. (2001). Syllables and morphemes: Contrasting frequency effects in Spanish. Journal of Experimental Psychology: Learning, Memory, and Cognition 27(2): 545-555.

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language* 40(5-6): 509-559.

Anderson, J.R., and Schooler, L.J. (1991). Reflections of the environment in memory. *Psychological Science* 2(6): 396-408.

Arnon, I., and Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition* 122(3): 292-305.

Arnon, I., and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62(1): 67-82.

Baayen, H. (1993). On frequency, transparency and productivity. In: *Yearbook of Morphology 1992* (eds. G. Booij and J. van Marle), pp. 181-208. Dordrecht: Springer.

Baayen, R.H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3): 436-461.

Baayen, R.H., Hendrix, P., and Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech* 56(3): 329-347.

Baayen, R.H., McQueen, J.M., Dijkstra, T., and Schreuder, R. (2003). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In: *Morphological structure in*

Language Processing (eds. R.H. Baayen and R. Schreuder), pp. 355-390. Berlin: Mouton De Gruyter.
Baayen, R.H., Milin, P., Filipović Đurđević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3): 438-481.
Baayen, R.H., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology* 30(11): 1174-1220.
Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., ... and Treiman, R. (2007). The English lexicon project. *Behavior Research Methods* 39(3): 445-459.
Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic.* Amsterdam: John Benjamins.
Beckner, C. (2018). The evidence add ups: An affix shift study of prefabs. In K. A. Smith & D. Nordquist (Eds.), *Functionalist and usage-based approaches to the study of language: In honor of Joan L. Bybee* (pp.199-224). Amsterdam: John Benjamins.
Bertram, R., Baayen, R.H., and Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language* 42(3): 390-405.
Bilalić, M., McLeod, P., and Gobet, F. (2008). Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology* 56(2): 73-102.
Blumenthal-Dramé, A., Glauche, V., Bormann, T., Weiller, C., Musso, M., & Kortmann, B. (2017). Frequency and chunking in derived words: A parametric fMRI study. *Journal of Cognitive Neuroscience*, *29*(7), 1162-1177.
Boudelaa, S., and Gaskell, M.G. (2002). A re-examination of the default system for Arabic plurals. *Language and Cognitive Processes* 17(3): 321-343.
Brothers, T., and Kuperberg, G.R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language* 116: 104174
Bryan, W. L., and Harter, N. (1899). The acquisition of a hierarchy of habits. *Psychological Review* 6: 345-375.
Brysbaert, M., and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4): 977-990.
Burke, D.M., Locantore, J.K., Austin, A.A., and Chae, B. (2004). Cherry pit primes Brad Pitt: Homophone priming effects on young and older adults' production of proper names. *Psychological Science* 15(3): 164-170.
Buz, E., Tanenhaus, M.K., and Jaeger, T.F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language* 89: 68-86.
Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.
Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5): 425-455.
Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.
Bybee, J. (2002a). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14(3): 261-290.

Bybee, J. (2002b). Sequentiality as the basis of constituent structure. In: *The Evolution of Language out of Pre-language* (eds. T. Givon and B. F. Malle), pp.109-134. Amsterdam: John Benjamins.
Bybee, J. (2003). Cognitive processes in grammaticalization. In: *The new psychology of language* (ed. M. Tomasello), pp. 151-174. Mahwah, NJ: Psychology Press.
Bybee, J.L., and Brewer, M.A. (1980). Explanation in morphophonemics: changes in Provençal and Spanish preterite forms. *Lingua* 52(3-4): 201-242.
Bybee, J.L., Chakraborti, P., Jung, D., and Scheibman, J. (1998). Prosody and segmental effect some paths of evolution for word stress. *Studies in Language* 22(2): 267-314.
Bybee, J., and Easterday, S. (2019). Consonant strengthening: A crosslinguistic survey and articulatory proposal. *Linguistic Typology* 23(2): 263-302.
Bybee, J., and Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2): 323-355.
Bybee, J.L., and Hopper, P.J., eds. (2001). *Frequency and the emergence of linguistic structure*. Amstrdam: John Benjamins.
Bybee, J., and Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics* 37(4): 575-596.
Carroll, R., Svare, R., and Salmons, J.C. (2012). Quantifying the evolutionary dynamics of German verbs. *Journal of Historical Linguistics* 2(2): 153-172.
Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22(6): 991-1013.
Clahsen, H., Rothweiler, M., Woest, A., and Marcus, G.F. (1992). Regular and irregular inflection in the acquisition of German noun plurals. *Cognition* 45(3): 225-255.
Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language* 93(3): 569-597
Corkery, M., Matusevych, Y., and Goldwater, S. (2019). Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. *arXiv preprint arXiv:1906.01280*.
Daelemans, W., and Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.
De Jong, N.H., Schreuder, R. and Baayen, R.H. (2000). The morphological family size effect and morphology, *Language and Cognitive Processes* 15: 329–36.
De Smet, H. (2016). How gradual change progresses: The interaction between convention and innovation. *Language Variation and Change* 28(1): 83-102.
Du, Y., Krakauer, J.W., and Haith, A. (2021). The relationship between habits and skills in humans. Preprint. https://doi.org/10.31234/osf.io/9qrgd
Ellis, N.C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics* 27(1): 1-24.
Ernestus, M., and Baayen, R.H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79(1): 5-38.
Ferreira, V.S., and Griffin, Z.M. (2003). Phonological influences on lexical (mis) selection. *Psychological Science* 14(1): 86-90.
Foster, K.I. (1976). Accessing the mental lexicon. In: *New approaches to language mechanisms* (eds. R. J. Wales and E. Walker), pp.257-287. Amsterdam: North-Holland.
Gahl, S. (2008). *Time* and *thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84(3): 474-496.

Giroux, I., and Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science* 33(2): 260-272.
Goldberg, A.E., Casenhiser, D.M., and Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics* 15(3): 289-316.
Goldiamond, I., and Hawkins, W.F. (1958). Vexierversuch: the log relationship between word-frequency and recognition obtained in the absence of stimulus words. *Journal of Experimental Psychology* 56(6): 457-463.
Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review 105*(2): 251-279.
Gollan, T.H., Montoya, R.I., Cera, C., and Sandoval, T.C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language* 58(3): 787-814.
Goodkind, A., and Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. Preprint. *arXiv preprint arXiv:2103.04469*.
Grainger, J., O'Regan, J.K., Jacobs, A.M., and Segui, J. (1989). On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Perception and Psychophysics* 45(3): 189-195.
Greenberg, S.N., Healy, A.F., Koriat, A., and Kreiner, H. (2004). The GO model: A reconsideration of the role of structural units in guiding and organizing text online. *Psychonomic Bulletin and Review* 11(3): 428-433.
Gries, S.T. (2013). 50-something years of work on collocations: What is or should be next…. *International Journal of Corpus Linguistics* 18(1): 137-166.
Gries, S.T. (2012). Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language* 36(3): 477-510
Griffiths, T.L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J.B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14(8): 357-364.
Hare, M., Elman, J.L., and Daugherty, K.G. (1995). Default generalisation in connectionist networks. *Language and Cognitive Processes* 10(6): 601-630.
Harmon, Z., and Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological Review*.
Harmon, Z., and Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology* 98: 22-44.
Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics* 39(6): 1041-1070.
Hay, J. B., Pierrehumbert, J.B., Walker, A.J., and LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition* 139: 83-91.
Hayes, B., Siptár, P., Zuraw, K., and Londe, Z. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4): 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(2): 235-242.
Healy, A.F. (1994). Letter detection: A window to unitization and other cognitive processes in reading text. *Psychonomic Bulletin and Review* 1(3): 333-344.

Hoeffner, J.H., and McClelland, J.L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in development dysphasia? A computational investigation. *Proceedings of the 25th Annual Child Language Research Forum* (pp. 38-49).
Hoppe, D.B., Hendriks, P., Ramscar, M., and van Rij, J. (2020). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective. Preprint. https://psyarxiv.com/py5kd
Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *The Journal of the Acoustical Society of America* 29(2): 296-305.
Howes, D.H., and Solomon, R.L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology* 41(6): 401-410.
Hsu, A.S., and Griffiths, T.E. (2010). Effects of generative and discriminative learning on use of category variability. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp.242-247). Austin, TX: Cognitive Science Society.
Jamieson, R. K., Crump, M. J., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior*, *40*(1), 61-82.
Johns, B.T., and Jones, M.N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology* 69(3): 233-251.
Jones, M.N., Johns, B.T., and Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology* 66(2): 115-124.
Juhasz, B.J., Yap, M.J., Raoul, A., and Kaye, M. (2019). A further examination of word frequency and age-of-acquisition effects in English lexical decision task performance: The role of frequency trajectory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45(1): 82-96.
Jurafsky, D., Bell, A., Gregory, M., and Raymond, W.D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Bybee and Hopper (2001), pp.229-254.
Kao, S.F., and Wasserman, E.A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19(6): 1363-1386.
Kapatsinski, V. (2006). Towards a single-mechanism account of frequency effects. *LACUS Forum* (Vol. 32, pp. 325-335).
Kapatsinski, V. (2009). Adversative conjunction choice in Russian (no, da, odnako): Semantic and syntactic influences on lexical selection. *Language Variation and Change* 21(2): 157-173
Kapatsinski, V. (2010a). Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech* 53(1): 71-105.
Kapatsinski, V. (2010b). 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(2): 157-215.
Kapatsinski, V. (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language* 89(1): 110-148.
Kapatsinski, V. (2014). What is grammar like? A usage-based constructionist perspective. *Linguistic Issues in Language Technology* 11: 1-41.

Kapatsinski, V. (2018a). *Changing minds changing tools: From learning theory to language acquisition to language change*. Cambridge, MA: MIT Press.
Kapatsinski, V. (2018b). Words versus rules (Storage versus online production/processing) in morphology. *Oxford Research Encyclopedia of Linguistics* (ed. M. Aronoff).
Kapatsinski, V. (2021a). Hierarchical inference in sound change: Words, sounds, and frequency of use. *Frontiers in Psychology* 12.
Kapatsinski, V. (2021b). Learning fast while avoiding spurious excitement and overcoming cue competition requires setting unachievable goals: reasons for using the logistic activation function in learning to predict categorical outcomes. *Language, Cognition and Neuroscience* 1-22.
Kapatsinski, V., Easterday, S., and Bybee, J. (2020). Vowel reduction: A usage-based perspective. *Rivista di Linguistica* 32: 19-44.
Kapatsinski, V., and Harmon, Z. (2017). A Hebbian account of entrenchment and (over)-extension in language learning. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol.39, 2366-2371). Austin, TX: Cognitive Science Society.
Kapatsinski, V., and Radicke, J. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In *Formulaic Language*, *volume 2: Acquisition, loss, psychological reality* (eds. Corrigan, R., Moravcsik, E.A., Ouali, H., and Wheatley, K.), pp.499-520. Amsterdam: John Benjamins.
Köpcke, K.-M. (1988). Schemas in German plural formation. *Lingua* 74: 303-335.
Köpcke, K.-M. (1998). The acquisition of plural marking in English and German revisited: schemata versus rules. *Journal of Child Language* 25(2): 293-319.
Koranda, M., Zettersten, M., and MacDonald, M. (2021). Good-enough production: Selecting easier words instead of more accurate ones. Preprint. https://psyarxiv.com/q2h9d/
Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.
Landauer, T.K., and Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior* 12(2): 119-131.
Liu, L., and Hulden, M. (2021). Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models. Preprint. *arXiv:2104.06483*.
Lohmann, A. (2018). Time and thyme are not homophones: A closer look at Gahl's work on the lemma-frequency effect, including a reanalysis. *Language* 94(2): e180-e190.
Luce, P. A., and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19(1): 1-32.
Madlener, K. (2016). Input optimization. In *Experience Counts: Frequency Effects in Language* (eds. H. Behrens and S. Pfaender), pp.133-174. Berlin: De Gruyter.
Marcus, G.F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology* 29(3): 189-256.
McCauley, S.M., Bannard, C., Theakston, A., Davis, M., Cameron-Faulkner, T., and Ambridge, B. (2021). Multiword units lead to errors of commission in children's spontaneous production: "What corpus data can tell us?". *Developmental Science*.
McCauley, S.M., and Christiansen, M.H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review* 126(1): 1–51.

McClelland, J.L., Botvinick, M.M., Noelle, D.C., Plaut, D.C., Rogers, T.T., Seidenberg, M.S., and Smith, L.B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences* 14(8): 348-356.
McClelland, J.L., and Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review* 88(5): 375-407.
McCurdy, K., Goldwater, S., and Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. Preprint. *arXiv:2005.08826*.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207-238.
Moder, C.L. (1992). *Productivity and categorization in morphological classes*. (PhD thesis, SUNY Buffalo).
Morrison, C.M., Ellis, A.W., and Quinlan, P.T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory and Cognition* 20(6): 705-714.
Morton, J. (1969). Interaction of information in word recognition. *Psychological Review* 76(2): 165-178.
Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, R.H. (2004a). Morphological family size in a morphologically rich language: the case of Finnish compared with Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30(6): 1271-1278.
Moscoso del Prado Martín, F., Ernestus, M., and Baayen, R.H. (2004a). Do type and token effects reflect different mechanisms? Connectionist modeling of Dutch past-tense formation and final devoicing. *Brain and Language* 90(1-3): 287-298.
Moscoso del Prado Martín, F., Kostić, A., and Baayen, R.H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94(1): 1-18.
Mowrey, R., and Pagliuca, W. (1995). The reductive character of articulatory evolution. *Rivista di Linguistica* 7: 37-124.
Murray, W.S., and Forster, K.I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review* 111(3): 721-756.
Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review* 113(2): 327-357.
Norris, D., and McQueen, J.M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review* 115(2): 357-395.
Nosofsky, R.M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(1): 54-65.
O'Donnell, T.J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. Cambridge, MA: MIT Press.
Oldfield, R. C., and Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17(4): 273-281.
Olejarczuk, P., and Kapatsinski, V. (2018). The metrical parse is guided by gradient phonotactics. *Phonology* 35(3): 367-405.
Olejarczuk, P., Kapatsinski, V., and Baayen, R.H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard* 4(s2).
Perfors, A., Ransom, K., and Navarro, D. (2014). People ignore token frequency when deciding how widely to generalize. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, 2759-2764). Austin, TX: Cognitive Science Society.

Phillips, B.S. (1984). Word frequency and the actuation of sound change. *Language* 60(2): 320-342.
Phillips, B.S. (2001). Lexical diffusion, lexical frequency, and lexical analysis. In Bybee and Hopper (2001), pp.123-136.
Pierrehumbert, J. (2001). Word frequency, lenition and contrast. In Bybee and Hopper (2001), pp.137-157.
Pinker, S. (1999). *Words and rules: The ingredients of language*. New York, NY: Basic Books.
Plaster, K., and Polinsky, M. (2010). Features in categorization, or a new look at an old problem. In *Features: Perspectives on a Key Notion in Linguistics* (eds. Kibort, A., and Corbett, G.G.). Oxford: Oxford University Press.
Plaut, D.C., and Booth, J.R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review* 107(4): 786-823.
Rácz, P., Beckner, C., Hay, J.B., and Pierrehumbert, J.B. (2020). Morphological convergence as on-line lexical analogy. *Language* 96(4): 735-770.
Raymond, W.D., and Brown, E.L. (2012). Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In *Frequency Effects in Language Learning and Processing* (eds. Gries, S.T., and Divjak,D.), pp.35-52. Berlin: Mouton De Gruyter.
Rogers, T.T., and McClelland, J.L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
Schneider, U. (2020). ΔP as a measure of collocation strength. *Corpus Linguistics and Linguistic Theory* 16(2): 249-274.
Schreuder, R., and Baayen, R.H. (1997). How complex simplex words can be. *Journal of Memory and Language* 37(1): 118-139.
Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133: 140-155.
Skousen, R. (1989). *Analogical modeling of language*. Dordrecht: Springer.
Sóskuthy, M., and Hay, J. (2017). Changing word usage predicts changing word durations in New Zealand English. *Cognition* 166: 298-313.
Sosnik, R., Hauptmann, B., Karni, A., and Flash, T. (2004). When practice leads to co-articulation: the evolution of geometrically defined movement primitives. *Experimental Brain Research* 156(4): 422-438.
Srinivasan, M., and Winter, B. (2021). Why is semantic change asymmetric? The role of concreteness and word frequency and metaphor and metonymy. Preprint. https://doi.org/10.31234/osf.io/zkycw
Stefanowitsch, A., and Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-243.
Suttle, L., and Goldberg, A. E. (2011). The partial productivity of constructions as induction. *Linguistics* 49(6): 1237-1269.
Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology Section A 57*(4): 745-765.
Tantucci, V., and Di Cristofaro, M. (2020). Entrenchment inhibition: Constructional change and repetitive behaviour can be in competition with large-scale "recompositional" creativity. *Corpus Linguistics and Linguistic Theory* 16(3): 547-579.

Theakston, A.L. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development* 19(1): 15-34.
Tiersma, P.M. (1982). Local and general markedness. *Language* 58(4): 832-849.
Tomaschek, F., Tucker, B.V., Fasiolo, M., and Baayen, R.H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard* 4(s2).
Tremblay, A., and Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In *Perspectives on formulaic language: Acquisition and communication* (ed. D. Wood), pp.151-173. London: Continuum.
Vogel Sosa, A., and MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language* 83(2): 227-236.
Wang, H. S., and Derwing, B. L. (1994). Some vowel schemas in three English morphological classes: Experimental evidence. In *In honor of Professor William S.-Y. Wang: Interdisciplinary studies on language and language change* (eds. Chen, M.Y., and Tang, O.C.L.), pp. 561-575. Taipei: Pyramid Press.
Xu, F., and Tenenbaum, J.B. (2007a). Word learning as Bayesian inference. *Psychological Review* 114(2): 245-272.
Xu, F., and Tenenbaum, J.B. (2007b). Sensitivity to sampling in Bayesian word learning. *Developmental Science* 10(3): 288-297.
Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. Cambridge, MA: MIT Press.
Zeldes, A. (2012). *Productivity in argument selection*. Berlin: Mouton De Gruyter.
Zipf, G.K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Reading, MA: Addison-Wesley.