Quantifying availability through linguistic productivity

Kyle Mahowald (kylemaho@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu)
Timothy J. O'Donnell (timod@mit.edu)
Brain and Cognitive Sciences, MIT

August 7, 2014

Keywords: Language, Availability, Memory, Productivity, Morphology

Corresponding author:
Kyle Mahowald
kylemaho@mit.edu
77 Massachusetts Ave, Bldg 46-3037
Cambridge, MA 02139
617.258.9344

Abstract:

In a classic study, Tversky and Kahneman showed that people guess that word patterns that include an English suffix (e.g., _ _ _ _ i ng) occur with a greater number of English words than patterns consisting of non-linguistic sub-sequences, (e.g., _ _ _ _ _n _), despite the fact that any word matching the former must necessarily match the latter. This result has been attributed to availability: the ease with which different representations can be retrieved from memory. Historically, availability has been a difficult concept to study quantitatively due to lack of suitable formalizations. Here we formalize and quantify the availability of linguistic units, such as suffixes, using the notion of productivity: how readily a unit can be combined with a base to form a novel word (e.g., *pine-scented*/*pine-scentedness*). In a large-scale behavioral experiment, we find that people systematically overestimate the frequency of word patterns that contain English suffixes and that the rate of overestimation is related to the suffix's productivity such that highly generalizable suffixes (like -*ness*) lead to more overestimation than less productive suffixes (like –*th*; *warmth*).

1. Introduction

Tversky and Kahneman famously claimed that people use the *availability heuristic* to reason about the probability of events, relying on the ease with which events can be retrieved from memory to estimate their probability, rather than using more veridical estimation mechanisms. An example of this phenomenon in language comes from a classic study in which participants judged (partial) word patterns that included a full English suffix (e.g., _ _ _ _ i n g) to be consistent with a greater number of English words than patterns that included a subset of letters from the suffix, (e.g., _ _ _ _ _n _), despite the fact that every word that matches the former pattern necessarily matches the latter pattern (Tversky & Kahneman, 1973, 1983). Tversky and Kahneman claim this failure occurs because linguistic memory is organized in such a way to make complete morphemes like *-ing* easier to access than arbitrary letter sequences like -*_n_*. In this paper, we generalize Tversky and Kahneman's empirical results, and propose a formal, quantitative account of overestimation rates in the special case of full suffixes.

Tversky and Kahneman's results reveal a fundamental property of human memory: It is organized around ecologically relevant units. In the case of language, these consist of the basic, mentally stored building blocks of utterances, such as words and affixes. This fact explains the greater availability of suffixes like *-ing* compared to non-linguistic sequences like –*_n_*. However, the problem of quantifying degrees of availability still remains. Why do people overestimate the probability of *-ing* as much as they do? One possibility is that this reveals another property of memory: Memory is not just a simple device for tracking experience but a tool for generalizing about the future (Anderson and Milson, 1989; Bartlett, 1932; Schacter, Addis, and Buckner, 2007). Using language requires being able to

understand words and phrases that have never been spoken or uttered before and, therefore, linguistic memory should be organized around units that often take part in the generation of new expressions. Presented with a pattern containing a full English suffix, participants will consider not only existing words containing the suffix, but also possible, but non-existing words that contain the suffix (a potentially unbounded number). The high rate of overestimation for *–ing* is predicted because *–ing* is a suffix that can be easily used to form novel words.

In linguistics, the ability of a rule, or structure-building process to create novel forms is known as its *productivity* (see, e.g., Bauer, 2001, 2005; Plag, 2004; O'Donnell, 2011, for reviews). In this paper, we explore the hypothesis that the rate at which people overestimate full suffixes like *-ing* depends on their productivity. Note that we do not question the empirical status or logic of Tversky and Kahneman's original result that full linguistic units such *–ing* are more available than non-linguistic sequences, like *-_n_*. Here we focus instead on the question of whether rates of generalization among linguistic units affect availability, comparing overestimation rates of **full** suffixes with differing levels of productivity.

English suffixes vary widely in their productivity. For example, the suffix *-ness* can be readily be used to form novel words (e.g., *Lady Gaga-esque-ness*). The suffix *-ity* is less productive: Although *-ity* does appear in a large number of existing words (e.g., *scarcity, sparsity*), most novel uses of the suffix (e.g., *coolity*) are impossible in English. However, *-ity* can be generalized in certain contexts, for instance, after the suffix *–able* (e.g., *googleability*). By contrast, the suffix *–th* (e.g., *width*, *warmth*) cannot be generalized to create new words at all in modern English. We predict that, because of its central role in making predictions about future word occurrences, differences in productivity such as those just outlined will affect

rates of overestimation for full English suffixes in experimental paradigm of Tversky and Kahneman.

This prediction regarding linguistic structure is a special case of more general, *constructive* accounts of memory organization, which have been especially prevalent in the rational analysis tradition (e.g., Anderson and Milson, 1989; Anderson, 1990; Anderson and Schooler, 1991; Bartlett, 1932; Huttenlocher et al. 1991, 2000; McClelland and Chappell, 1998; Schacter, 2012; Schacter, Addis, and Buckner, 2007; Shiffrin and Steyvers, 1997; Hemmer and Steyvers, 2009; Steyvers and Hemmer, 2012). It is well known that memory is influenced by general knowledge about the content and context of memories, relying on both specific (although noisy) information about stored items and general expectations about the internal structure or categories of those items. A variety of experimental tasks, examining both episodic and semantic memory, have demonstrated that people "misremember" unstudied items that share properties with studied items, effectively generalizing along certain dimensions of stimulus structure. For example, when asked to remember the values of continuous properties associated with specific items, such as size, people tend to misremember values which are closer to the mean of the category from which the stimuli were drawn (Huttenlocher et al. 1991, 2000; Hemmer and Steyvers, 2009; Steyvers and Hemmer, 2012). Similarly, when recalling memorized words, people will often falsely remember semantically related words (e.g., Roediger and McDermott, 1995). In constructive theories of memory, such data are usually accounted for by assuming that processes of encoding and/or retrieval combine veridical information from studied exemplars with more general prior expectations about categories. Under this view, generalization is a design-feature of the system because it supports category-based inferences from noisy input.

The mental storage of words can be viewed from a similar perspective. Some words forms are stored veridically as whole-forms in memory (e.g., *dog*, *warmth*) and some are built from other stored parts on the fly (e.g., *pine-scentedness* = *pine-scented* + *-ness*). It is well established that word comprehension and production consist of a complex mixture of both whole-form retrieval and composition from parts (see, Hay, 2003; O'Donnell, 2011, for reviews). Because the set of words which can be formed using a word-part like *–ness* can be viewed as a category which admits the generalization of previously unseen word forms (see, e.g., Aronoff, 1976; Baayen, R. H., 1992), memory structures underlying word-processing can be seen as implementing a mixture of veridical information and category-based generalization similar to rational accounts of memory in other domains. Differences in productivity between suffixes such as *–ness* and *–th* reflect differences in the readiness with which word categories generalize.

To test the relative effects of productivity, we extend the experimental paradigm of Tversky and Kahneman (1983), obtaining empirical frequency estimates for a variety of word frames which including a subset containing a large number of full English (derivational) suffixes. We first replicate the original Tversky and Kahneman results, showing that word frames that contain an English suffix are systematically overestimated in comparison to word frames that do not contain a whole suffix. We then run a second experiment focusing just on frames with full English suffixes, predicting that more highly productive suffixes will show greater rates of overestimation than less productive suffixes. We quantify productivity using three leading proposals from the literature: a Bayesian generative model (Fragment Grammars; O'Donnell,

2011), a Good-Turing based estimator (Baayen's $\mathcal{P}^*$ Baayen, 1994), and an estimator based on type frequency of the suffix (log type frequency of each suffix; Bybee, 1995). We find, as predicted, that higher productivity suffixes predict greater rates of frequency overestimation for all three predictors.

2. Experiment 1

2.1 Methods

Materials

Each participant saw a sample of 105 word frames drawn from the following four categories: (i) full-suffix frames like `_ _ _ _ n e s s`, (ii) partial-suffix frames like `_ _ _ _ n _ s _`, (iii) frames based on mono- morphemic words like `r _ _ d` (*road*, *reed*, etc.) that act as filler items, and (iv) impossible frames like `_ _ _ o a e` (used as catch trials to prevent random guessing).

Full-suffix trials were pseudo-randomly sampled from 75 unique suffixes drawn from the database of morphologically complex English words constructed by O'Donnell (2011). Of these 75 suffixes, each participant saw 25 full suffix frames and 25 partial suffix frames. Partial suffixes were created by randomly deleting letters from the full suffix (for example, for *–esque*: `_s q u e`, `_ _ q _ e`). Since longer suffixes like *–esque* have a greater number of partial suffixes than shorter suffixes, partial suffixes which were created from longer suffixes were sampled more often than ones formed from shorter suffixes so as to sample a range of possible deletions from longer suffixes.

Frames were created for both partial and full suffixes by concatenating them with an empty

"stem" (e.g., _ _ _ _) whose length was either (i) the mean stem length for that suffix, (ii) the mean stem length for that suffix minus one standard deviation, or the mean stem length for that suffix plus one standard deviation (rounded to the nearest integer). For example, in the full suffix condition, *-ness*, whose mean stem length was 6 with a standard deviation of 2, was equally likely to appear with an empty 4 letter stem (i.e., _ _ _ _ n e s s), 6 letter stem (i.e., _ _ _ _ _ _ n e s s), or 8 letter stem (i.e.,

_ _ _ _ _ _ _ _ n e s s).

Mono-morphemic frames, of which each participant saw 30, were randomly chosen with the aim of presenting a wide variety of frames, from those with many possible completions like s _ _ _ _ to those with very few like b r i c _. Mono-morphemic frames were uniquely generated for each participant.

Impossible frames, of which each participant saw 25, were created by taking the full and partial frames, randomly replacing the existing letters with new ones, and checking to make sure that there were no words in SUBTLEX that fit that pattern.

Participants

Using Amazon's Mechanical Turk, we presented 225 participants (whose IP addresses were restricted to the United States) with surveys in which they estimated the frequency of word frames with one or more missing letters (e.g., _ r _ _). Due to processing errors, only 223 surveys were collected. 206 participants remained after excluding self-identified non-native English speakers, participants who took the survey more than once, participants who failed to provide answers for more than 90% of trials, and participants who gave higher mean estimates for impossible trials than for one or more of the other conditions. All participants

were able to complete the task in well under the allotted time.

Procedure

Participants were given the following directions.

Imagine that you just read a modern novel that was about 100,000 words long. In those 100,000 words, how many of those words do you think fit the given pattern? If the pattern were _ _ r _ _ (i.e., a 5 letter word whose third letter is "r"), a good guess might be 1,000 words. If the pattern were _ _ u l _ (a five letter word whose middle letters are "ul"), a good guess would be 370. For some patterns, there are no possible words that fit that pattern, and the answer will be 0. It is unlikely that any given pattern will have more than 10,000 matches in the 100,000 words. So guesses over 10,000 are typically not good estimates. Because some Turkers randomly guess, we will be checking to make sure that estimates are not the product of random guessing. Do not consult a dictionary or any other resources. We are interested in your intuition.

2.2 Results

We first tested whether our results replicated Tversky and Kahnemans's original result by analyzing how estimates varied across full suffix frames, partial suffix frames, and mono-morphemic frames. We then looked specifically at the variation within the full suffix frames to see if there were differential effects of productivity on estimation rates when five factors were controlled: (i) the actual token frequency of the frame (i.e., the total frequency of words in SUBTLEX which were consistent with the frame), (ii) the type frequency of the frame (i.e., the number of words in SUBTLEX which were consistent with the frame), (iii) the number of letters present in the frame, (iv) the number of letters missing from the frame, and (v) the interaction between the number of letters present and the number of letters missing.

We restricted our analyses to data from suffixes of 2 or more letters and those suffixes for which we have productivity predictions (from O'Donnell, 2011). This left 51 suffixes out of the original 75.

Consistent with the findings of Tversky and Kahneman, the median estimate for mono-morphemic frames was 3.14 (95% CI of the sample median by non-parametric bootstrap [2.92, 3.34]) times greater than the actual estimated frequency based on SUBTLEX (Brysbaert & New, 2009), whereas it was 7.65 [95% CI 7.04, 8.49] times greater than the SUBTLEX estimates for partial-suffix frames and 29.53 [95% CI 27.38, 31.22] times greater for full-suffix frames.

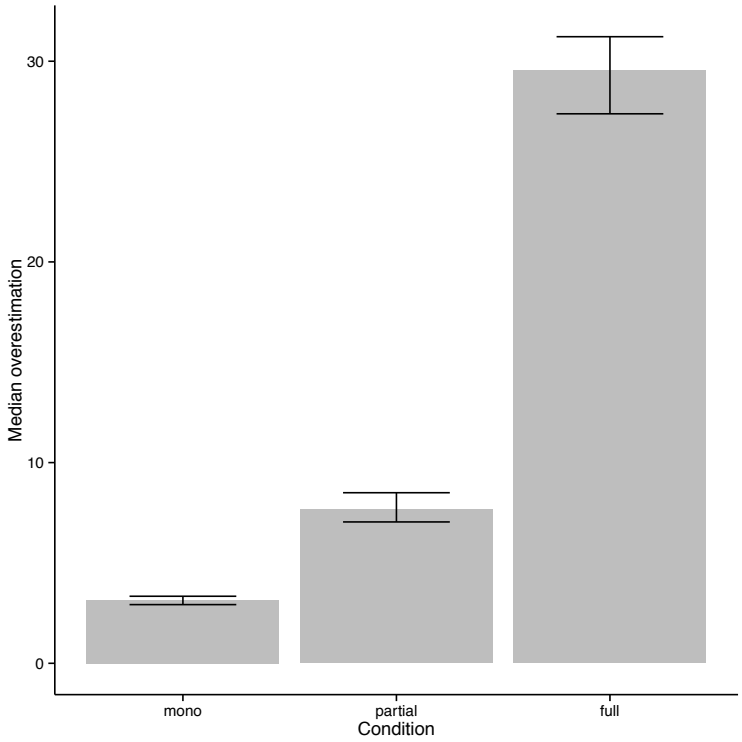

Figure 1 Figure 1 shows the median ratio of estimated frequency to actual token frequency for each type of frame (full suffix, partial suffix, and mono-morphemic). 95% confidence intervals for each median are estimated using a bootstrap (i.e. sampling with replacement and calculating the median).

To test for an effect of suffix status in this dataset, after excluding catch trial, we fit a mixed effect model predicting the estimated frequency of each frame, while controlling

for (i) the log token frequencies of the frame (i.e., the total frequency of all words consistent with the frame from SUBTLEX), (ii) the log type frequency of the frame (i.e., the number of words consistent with the frame from SUBTLEX), (iii) the number of letters present in the frame (iv) the number missing letters in the frame, and (v) the interaction between present and missing letters (centered predictors and a maximal random effect structure by subject with +1 smoothing to avoid log error). Using Helmert-coded predictors for, the effect of having a suffix (pooling across full-suffix frames and partial-suffix frames) was large: people significantly overestimated the frequency of full and partial suffix frames relative to mono-morphemic frames ($\beta = .44$, $t = 13.36$, $p < .001$). Moreover, replicating the Tversky and Kahneman result, people overestimated the frequency of full suffix frames relative to partial suffix frames ($\beta = .24$, $t = 6.57$, $p < .001$).

3. Experiment 2: Effect of productivity on full suffixes

3.1 Methods

Having established the main effect of inflated frequency estimates for frames with full suffixes, we then asked whether there is an additional effect of productivity of withing the set of full suffixes. We predicted that more productive suffixes should receive inflated estimates relative to less productive suffixes. In this section, using the experimental paradigm and analysis described above, we tested a new pool of subjects on only the 40 full-suffix frames (frames like _ _ _ _ n e s s and _ _ _ _ i t y) that contained than 2 characters and for which we had productivity predictions available.

Materials

We presented subjects with 40 full suffix frames. Each suffix could be seen with each of 3 possible stem lengths, as described in Experiment 1, and each subject saw each suffix once. There were 5 catch trials (obviously impossible frames like `_ _ _ _ q v x`) but otherwise no filler items.

Participants

We presented the survey to 240 subjects recruited on Amazon's Mechanical Turk so that each suffix was seen 240 times, 80 times each with each of its 3 possible stem lengths. 231 participants remained after excluding self-identified non-native English speakers, participants who took the survey more than once, participants who failed to provide answers for more than 90% of trials, and participants who gave higher mean estimates for impossible trials than for one or more of the other conditions.

Procedure

The instructions to participants were identical to those in Experiment 1.

3.2 Results

To assess the effect of productivity on people's estimates of these word frames, we derived productivity estimates using three models of productivity from the literature. The first, Fragment Grammars (O'Donnell, 2011), is a Hierarchal Bayesian generative model of lexical storage and computation. A Fragment Grammar acquires a lexicon of word and word-parts by finding an optimal balance between productivity and reuse in a particular training data set. By computing the probability of fragments associated with suffixes in the lexicon (e.g., the word-fragment that adds *-ness* to adjectives to form nouns), we can estimate the probability

that a particular suffix will give rise to novel word, that is, its productivity. The second measure of productivity we use, Baayen's $\mathcal{P}^*$ (Baayen, 1994), is an estimator of the conditional probability of an affix being used to form a new word, that is, P(-*suffix* | NOVEL). $\mathcal{P}^*$ draws on the Good-Turing theory of estimating unobserved events (Good, 1953), which states that estimates of unobserved events should be based on the number of events of the intended type that occur only once in a given sample. Thus, $\mathcal{P}^*$ is proportional to the number of words ending in a suffix that appear only one time in a corpus (the *hapax legomena*, in technical parlance). Finally, we also use the (log) number of distinct words using each suffix as an estimator of productivity. This quantity, known as the (log) type frequency,has been frequently proposed as a predictor of productivity in the literature (see, e.g., Bybee, 1995; Ambridge, et al., 2012).[1]

To test the effect of productivity on people's estimates of full word-frame frequencies, we fit a linear mixed effect model with the same controls as in Experiment 1 (log type and token frequency of the frame,[2] number of letters present in the frame, number of letters missing in the frame, and the interaction of those two terms). We then looked at the relationship between the residuals of this model and the productivity scores for each suffix (where productivity is estimated using one of the three measures discussed above), as shown in Figure 2.

[1] These three estimators sometimes produce divergent productivity predictions. For example, type frequency is a poor estimator of productivity when all tokens using some affix are also highly token-frequent (Baayen, 2006). Nevertheless, since these models have all been advocated in the literature, and since they produce similar results on this dataset, we include all three.

[2] Note that we controlled for the type frequency of the frame (i.e., the total number of distinct words in our corpus that matched the frame). This differs from the type frequency of the suffix (i.e., the total number of distinct words ending in a particular suffix), which we used as a predictor.

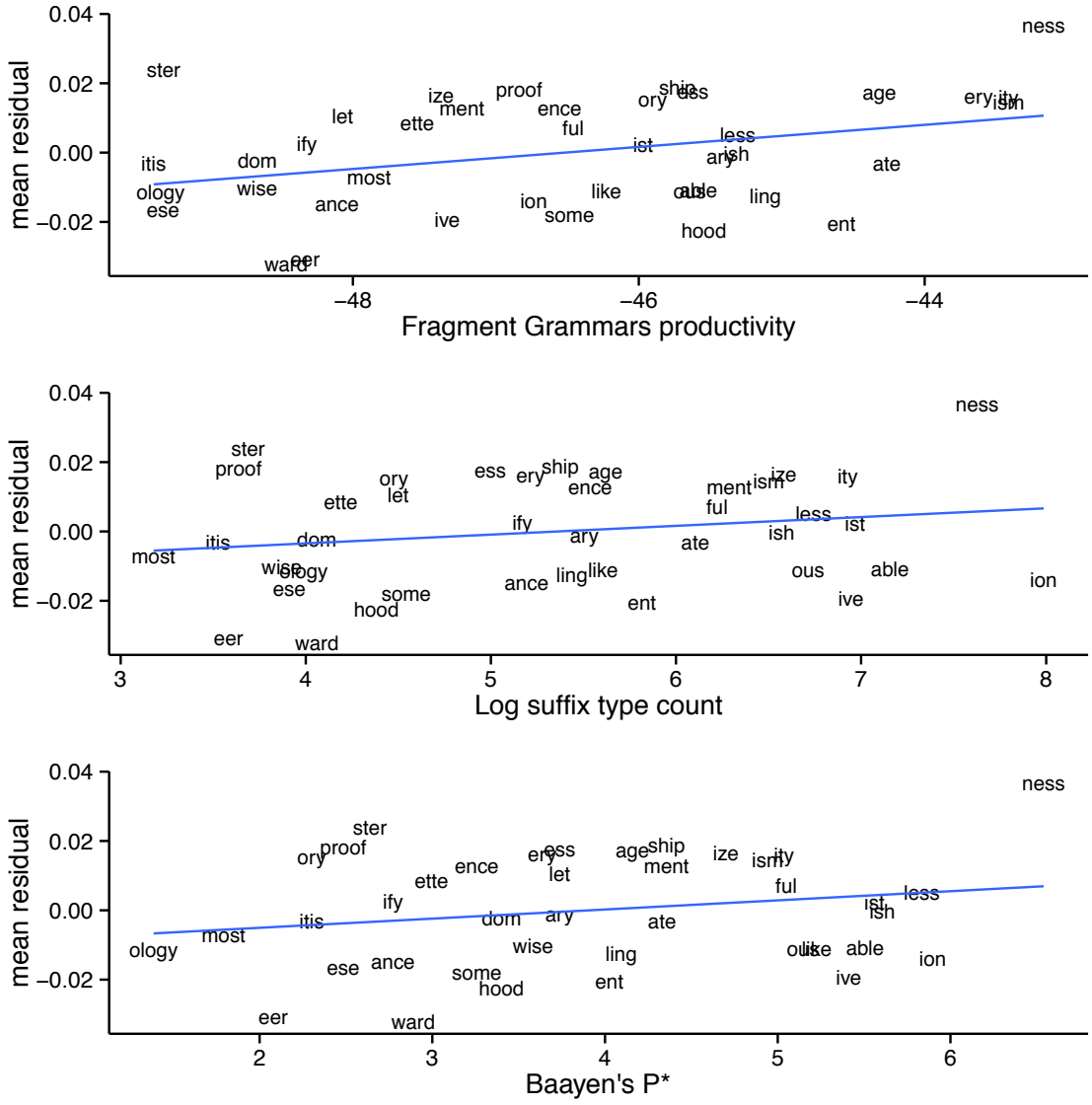

Figure 2 Each plot shows a different measure of productivity on the x-axis (Fragment Grammars productivity score, log type frequency of suffix, and Baayen's $\mathcal{P}$*) and the mean residuals from the model by suffix on the y-axis. The upward trend from left to right shows an effect of productivity on people's estimates. If there were no effect of productivity, we would expect to see no relationship between the x and y values.

If there were no effect of productivity above and beyond frequency and the other controls, we

would expect the residual plots to look like pure noise. Instead, we find an upward trend from left to right: all three productivity scores tested were predictive of the residuals. Probability of suffix generalization as estimated by Fragment Grammars (β=.042, t=6.56, $p < .0001$), Baayen's $\mathscr{P}^*$ (β=.064, t=7.10, $p < .0001$), and log type frequency of the suffix (β=.060, t=6.64, $p < .0001$) were all significantly predictive of the residuals. Thus, more productive suffixes have their frequency estimates systematically inflated relative to lower productivity suffixes.

We also find the same pattern of significant results in the results of Experiment 1 when we focus on just the full-suffix frames tested in Experiment 2. For Experiment 1, in a simple regression predicting residuals from the full mixed effect model as specified above for the Experiment 2 analysis, FG productivity score (β=.012, t=2.51, $p < .05$), log type frequency of suffix (β=.05, t=2.26, $p < .05$), and Baayen's P^* (β=.04, t=2.10, $p < .05$) are all significant predictors. Thus, the generalization replicates in two data sets with different participants.

4. Conclusion

Tversky and Kahnemans (1973, 1983) showed that people systematically overestimate the frequency of full suffixes, such as *–ing* compared to the frequency of non-linguistic sequences, such as - _ *n* _. This result has been attributed to the greater *availability* of linguistic representations during memory access. We have shown that this replicates for a much wider range of suffixes and non-linguistic sequences. Furthermore, we have shown that within full suffixes, an additional factor, linguistic *productivity*---the ability of a suffix to give rise to new forms---further explains differences in overestimation rates.

We emphasize that productivity can only have an effect on estimation rates once a participant

has identified a linguistically well-formed unit, such as *-ness*. The differences in availability originally identified by Tversky and Kahneman between linguistic sequences, such as *ing*, and non-linguistic sequences, such as *_ n_*, do not fall under the domain of our model. Predicting quantitative differences for these cases will likely require modeling additional, non-linguistic structure. We leave this to future work.

More broadly, we believe that this work supports the idea that generalization is a driving force behind memory. The role of productivity in these estimates suggests that overestimation of the frequency of full suffixes is likely caused by a "hallucinatory" effect of productivity on memory. More productive suffixes can give rise to a larger number of novel words, and this causes greater overestimation of frequency. Our results examining productivity in language—a domain in which quantitative models are available—suggest that productivity may play a role in other types of memory. Much as we can use a generative grammar for morphology to produce novel words, people might use grammars of visual episodes, causal relations, social roles, and other domains of cognition to remember things that were never there, but could be.

Acknowledgments

We gratefully acknowledge Sam Gershman for reading and providing detailed feedback on several drafts of this work, as well as Leon Bergen for detailed discussion and comments. We also thank Ted Gibson, members of Tedlab and Cocosci, and the audience at CUNY 2013 for helpful conversations.

References

Ambridge, B., Pine, J. M., Rowland, C., Chang, F., and Bidgood, A. (2012). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. WIREs Cognitive Science.

Anderson, J. R. and Milson, R. (1989). Human memory: An adaptive perspective. Psychological Review, 96(4):703–719.

Anderson, J. R. (1990). The Adaptive Character of Thought. Lawrence Erlbaum Associates, Hillsdale, NJ.

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

Aronoff, M. (1976). Word formation in generative grammar. Linguistic Inquiry Monographs. MIT Press, Cambridge, MA.

Baayen, R. H. (1994). Productivity in language production. *Language and Cognitive Processes*, *9*(3), 447–469.

Baayen, R. H. (2006). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook*. Mouton de Gruyter.

Bartlett, F. C. (1932). Remembering: A Study in Experimental and Social Psychology. Cambridge University Press.

Bauer, L. (2001). *Morphological productivity*. Cambridge: Cambridge University Press.

Bauer, L. (2005). Productivity: Theories. In *Handbook of word-formation*. Springer.

Brysbaert, M., & New, B. (2009, November). Moving beyond Kucera 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.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, *10*(5), 425–455.

Good, I. J. (1953, December). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3/4), 237–264.

Hay, J. (2003). *Causes and consequences of word structure*. New York, NY: Routledge.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Huttenlocher, J., Hedges, L. V., and Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. Psychological Review, 98(3):352–376.

Huttenlocher, J., Hedges, L. V., and Vevea, J. L. (2000). Why do categories affect stimulus judgment? Journal of Experimental Psychology: General, 129(2):220–241.

McClelland, J. L. and Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. Psychological Review, 105(4):724–760.

O'Donnell, T. J. (2011). *Productivity and reuse in language*. Unpublished doctoral dissertation, Harvard University.

Plag, I. (2004). Productivity. In *Encyclopedia of language and linguistics*. Elsevier.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Schacter, D.L. (2012). Adaptive constructive processes and the future of memory. *American Psychologist, 67*, 603-613.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience, 8*(9), 657-661.

Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. Psychological Bulletin and Review, 4(2):145–166.

Steyvers, M. and Hemmer, P. (2012). Reconstruction from memory in naturalistic environments. In The Psychology of Learning and Motivation. Elsevier.

Tversky, A., & Kahneman, D. (1973, September). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, *90*(4), 293.

Yang, C. (2005). On productivity. In Linguistic Variation Yearbook 5. John Benjamins.