Formalizing accessibility using 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

Abstract:
Tversky and Kahneman famously claimed that people use the *availability heuristic* to reason about the probability of events, relying on the ease with which information about the events can be retrieved from memory—a property known as *accessibility* (or sometimes *availability*)—to estimate probability, rather than using more veridical mechanisms. As an example of this phenomenon, Tversky and Kahneman showed that people guess that word patterns that include an English suffix (e.g., _ _ _ _ i n g) 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. T&K's explanation is that this discrepancy results because *-ing* is a linguistic unit---a suffix---and, therefore, more accessible in memory than the non-linguistic sequence - _ *n*_.. In this paper, we propose that the accessibility of linguistic units should be related to their *productivity*--how readily the units can be combined to form novel expressions (e.g., *pine-scented*/*pine-scentedness*). Since highly productive units, such as the suffix *–ing*, are frequently needed as independent units during the production and comprehension of novel words (e.g., *pine-scentedness*) they should be more accessible in memory than less productive suffixes, which are typically only accessed as parts of words (e.g., *-th* in *warmth*). We tested this idea in a large-scale behavioral experiment using T&K's paradigm with a variety of English suffixes of differing levels of productivity. We found that people systematically overestimated the frequency of word patterns that contained English suffixes and that the rate of overestimation was related to the suffix's productivity, such that highly generalizable suffixes (like -*ness*) led to more overestimation than less productive suffixes (like *–th*; *warmth*). This result provides support for the idea that (a) morphological productivity can be measured in behavioral paradigms and that (b) morphological productivity is connected to traditional notions of accessibility.

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 information about the events

can be retrieved from memory—a property known as *accessibility*[1]—to estimate 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 of estimation occurs because full suffixes like *–ing* serve as better retrieval cues than arbitrary letter sequences like -_*n*_. Since the concept was introduced by Tversky and Kahneman (1973), the availability heuristic has been used to explain a large number of phenomena across diverse areas, including event planning (Carroll, 1978), risk assessment (Shedler & Manis, 1986), stereotype formation (Rothbart, Fulero, Jensen, Howard, & Birrell, 1978), and many others (see Kahneman 2003 for a review).

But, what is accessibility? Kahneman (2003) defines accessibility as "the ease (or effort) with which particular mental contents come to mind." In Tversky and Kahneman's original experiment, the ease of bringing particular mental contents to mind was manipulated using the fact that *–ing* is an regularly used, generalizable linguistic suffix and, therefore, is (presumably) explicitly represented in memory, whereas there is no reason for the existence of a memory trace corresponding exactly to the pattern -_*n*_. The (non)existence of a mental representation of some category of stimuli is obviously a major determinant of accessibility.

Many determinants of accessibility have also been studied in the literature. One well-studied example is frequency: it is should be easier to mentally access a word you use everyday than a word that you use once every few years. But as Kahneman points out, the accessibility of a memory trace depends not only on raw frequency, but also on a complex set of other factors, such as the way in which the trace is represented, the characteristics of the stimuli which evoke it, and prior knowledge of the domain possessed by the thinker. Indeed, while a large number of determinants of accessibility have been studied in the literature (e.g., frequency, salience, selective attention, priming, associativity, and distinctiveness, *inter alia*), as Kahneman notes, "[…] much is known about the determinants of accessibility, but there is no general theoretical account of accessibility and no prospect of one emerging soon.

In this paper, we propose a formal theoretical account of the accessibility of existing linguistic units like *–ing* or *-ness*. In particular, we propose that the accessibility of linguistic units should be related to their *productivity*. Productivity refers to the ease with which a linguistic unit, such as the suffix *–ness*, to give rise to new forms, such as *pine-scentedness* and can be formalized as the probability that a particular unit will be used to form a new word (see, Bauer, 2001, 2005; Plag, 2004; Hay, 2003; O'Donnell, 2015, for discussion). Logically, if a suffix is used to produce or comprehend a completely novel word, it must

[1] Tversky and Kahneman originally called this property *availability*. However, in more modern work, the term accessibility is preferred (including by Kahneman; see Kahneman, 2003).

have some independent representation in the minds of speakers. Moreover, there is a great deal of independent evidence that productive affixes are decomposed during the processing of even frequent words (see, Aronoff 1976, 1978; Bauer, 2001, 2005; Plag, 2004; Hay, 2003; van Marle 1990; Marslen-Wilson 2007; O'Donnell, 2015, for reviews). Our proposal is that the accessibility of a suffix (as measured using T&K's experimental paradigm) should be directly related to the probability that the suffix is used to form novel words---it's productivity.

English derivational morphology exhibits suffixes that vary widely in their productivity, allowing us to test our hypothesis in that domain. For example, consider the three suffixes –*ness*, -*ity*, and –*th*. The suffix -*ness* can be readily be used to form novel words (e.g., *Lady Gaga/ Lady-Gaga-esque/Lady-Gaga-esque-ness*). The suffix –*ity*, which has similar meaning to –*ness*, is less productive: Although -*ity* does appear in a large number of existing words (e.g., *scarcity, sparsity*), in general, it cannot be used to form new words (cf., *coolity*). However, in certain contexts -*ity* does generalize, for example, after the suffix –*able* (e.g., *tweet*/*tweetable*/*tweetability*). By contrast, the suffix –*th* (e.g., *width*, *warmth*), which also has a similar meaning to –*ness* and -*ity*, cannot be generalized to create new words in modern English (cf., *coolth*) but, rather is only reused in small set of existing words. In this setting, the productivity of the suffix –*ness* is given by the probability that the memory structure [`Adj` –*ness*] will be used in the future. In other words, estimates of the productivity of the suffix –*ness* can be understood as estimates of the probability of needing to access the memory unit [`Adj` –*ness*]. Relatedly, when we say that a suffix such as –*th* has low productivity, we mean that it does not freely generalize to form new words, and existing words which contain it are rarely decomposed during processing. In other words, the memory structure, [`Adj` –*th*], has low probability of being accessed as an independent unit. In the extreme case, to say a suffix is completely unproductive is just to say that it has no independent combinatorial potential. This account could plausibly explain the asymmetry between –*ing* and -_*n*_ observed by Tversky and Kahneman: -*ing* has a frequently accessed memory unit [`Verb` –*ing*], whereas there is no such unit for -_*n*_. This account also provides predictions for differential accessibility between full suffixes such as –*ness*, –*ity* and –*th*.

To test our hypothesis that accessibility in natural language can be (at least partially) formalized using productivity, we investigated the degree to which three leading quantitative models of productivity are able to predict rates of frequency overestimation in T&K's experimental paradigm. Note that we do not question the empirical status or logic of Tversky and Kahneman's original result that full linguistic units such as –*ing* are more accessible than non-linguistic sequences, like -_*n*_. Instead, we focus on the empirical case for which explicit quantitative theories have been proposed: the productivity of full suffixes. We first replicated the original Tversky and Kahneman (1983) 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 ran 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 estimated the productivity of suffixes using three leading models from the literature: a Bayesian generative model (Fragment Grammars; O'Donnell, 2015), a Good-Turing based estimator (Baayen's $\mathscr{P}$* Baayen, 1994), and an estimator based on type frequency of the suffix (log type frequency of each suffix; Bybee, 1995). We found, as predicted, that more highly productive suffixes are associated with greater frequency overestimation. Our three models make highly correlated predictions for the set of suffixes we study in this paper, and thus show similar ability to predict overestimation rates.

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 (meaning any given participant saw 50 of the 75 possible suffixes). 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. Therefore, longer suffixes were somewhat more likely to be seen by any given participant than shorter ones.

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 (a database of lexical frequencies derived from U.S. movie subtitles and shown to have good correlation with behavioral correlates of frequency; see Brysbaert & New, 2009) 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 a technical error, 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, 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.

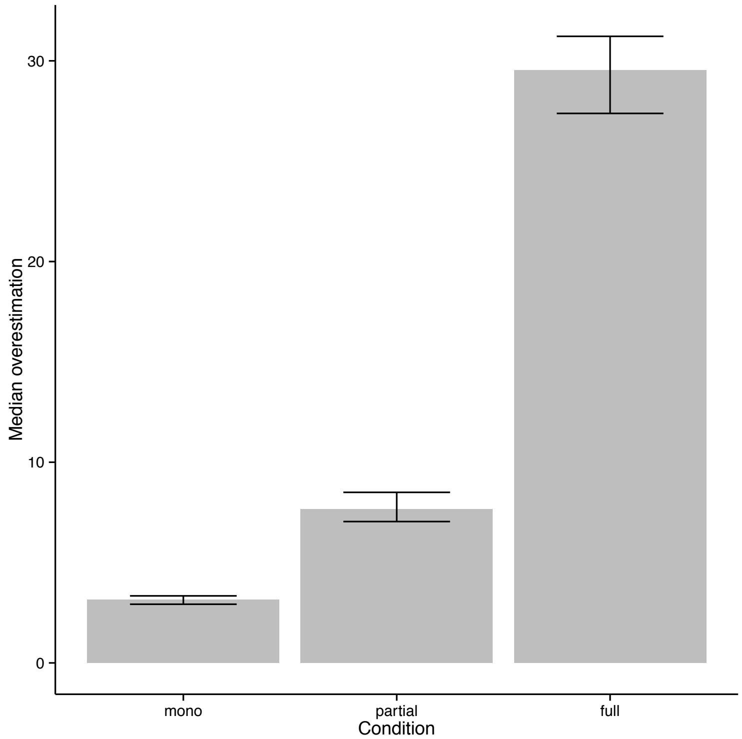

Figure 1 Figure 1 shows the median ratio of estimated frequency to actual token frequency for each type of frame (full suffix, partial suffix, and mono-morphemic). 95% confidence intervals for each median are estimated using a bootstrap (i.e. sampling with replacement and calculating the median).

To test for an effect of suffix status in this dataset, after excluding catch trial, we fit a mixed effect model using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) in R to predict 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 of missing letters in the frame, and (v) the interaction between present and missing letters (centered and scaled predictors and a maximal random effect structure by subject with +1 smoothing to avoid log error). Using Helmert-coded predictors, we found that (pooling across full-suffix frames and partial-suffix frames) people significantly overestimated the frequency of full and partial suffix frames relative to mono-morphemic frames (β = ..17, t = 13.4, $p < .0001$). The beta coefficient of .17 [95% CI .15, .19] indicates that, holding all else constant, the presence of a suffix predicts a .17 standard deviation increase in the log estimate. The presence of the Helmert-coded predictor for whether the frame has a suffix or is monomorphemic significantly improves the model by a likelihood ratio test comparing models with and without this predictor while holding all else, including random effects, constant ($\chi^2(1)$= 129, p < .0001). We also found the covariates --including token frequency of the frame ($\chi^2(1)$= 276.7, p < .0001), type frequency of the frame ($\chi^2(1)$= 115.2, p < .0001), and the structure of the frame as represented by the number of present letters, missing letters and their interaction ($\chi^2(3)$= 110.1, p < .0001)--to be highly significant for model fit in that the subtraction of any one of those fixed effect terms caused a significantly worse fit. Note that, even in the model including these highly predictive covariates, however, the inclusion of the morphological status of the frame significantly improved the model fit.

Moreover, replicating the Tversky and Kahneman result, people overestimated the frequency of full suffix frames relative to partial suffix frames (β = .09, t = 6.58 p < .0001). That is, having a full as opposed to partial suffix predits a .09 standard deviation increase in the log estimate given in the task. The presence of the Helmert-coded predictor for whether the frame has a full or partial suffix significantly improves the model by a likelihood ratio test comparing models with and without this fixed effect term while holding all other predictors and random effects constant (χ^2 (1) = 39.8, p < .0001).

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 within 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 more 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. These three models can all be understood as estimators for productivity. Because the estimates produced by these three models are correlated for the suffixes considered in this study, we simply include analyses of all three. However, we emphasize that in the general case, these models produce divergent predictions.

The first model, Fragment Grammars (O'Donnell, 2011), is a probabilistic 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 for 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 $\mathscr{P}^*$ (Baayen, 1994), is an estimator of the conditional probability of an affix being used to form a new word, that is, P(-*suffix* | `NOVEL`). $\mathscr{P}^*$ 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, $\mathscr{P}^*$ 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).[2]

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,[3] 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. 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.

We also fit linear mixed effect models using the covariates as described above with each of

---

[2] 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.

[3] 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.

the three models of productivity (because they are all highly correlated, we do not include all three models in the same regression) and a maximal random effect structure with intercepts by participant and by-participant slopes for all covariates. All predictors were scaled to have mean 0 and standard deviation 1. For the productivity score output from Fragment Grammars (which is in log space and then scaled), the model coefficient for the fixed effect term was .12 with a [.09, .15]. Thus, an increase in one standard deviation in productivity predicts a .12 standard deviation increase in the log of the experimental estimate given by participants for that frame. For Baayen's $\mathscr{P}^*$, the beta coefficient was .15 [.12, .18], meaning each standard deviation increase in Baayen's $\mathscr{P}^*$ predicts a .15 standard deviation increase in participants' estimate. And for the log number of suffix types, the beta coefficient was .18 [.16, .20]. Thus, for all 3 cases, the models of productivity appear to explain a meaningful portion of the variance in the model.[4] To assess significance formally, we then fit a model that does not include a fixed effect term for productivity but which is otherwise exactly the same and compared this model to the full model using a likelihood ratio test. For all three models of productivity, the productivity score significantly improves the model fit, as shown by a likelihood ratio test (for Fragment Grammars: χ^2 (1) = 53.7, p < .0001; for Baayen's $\mathscr{P}^*$: χ^2 (1) = 68.9, p < .0001; for log type frequency of the suffix: χ^2 (1) = 74.9, p < .0001). Thus, more productive suffixes have their frequency estimates systematically inflated relative to lower productivity suffixes.

[4] Because it is difficult to compute values for mixed effect models like this, particularly ones that have many correlated covariates, we focus on the beta values from the model and the uncertainty on those values. See Gelman & Hill (2007).

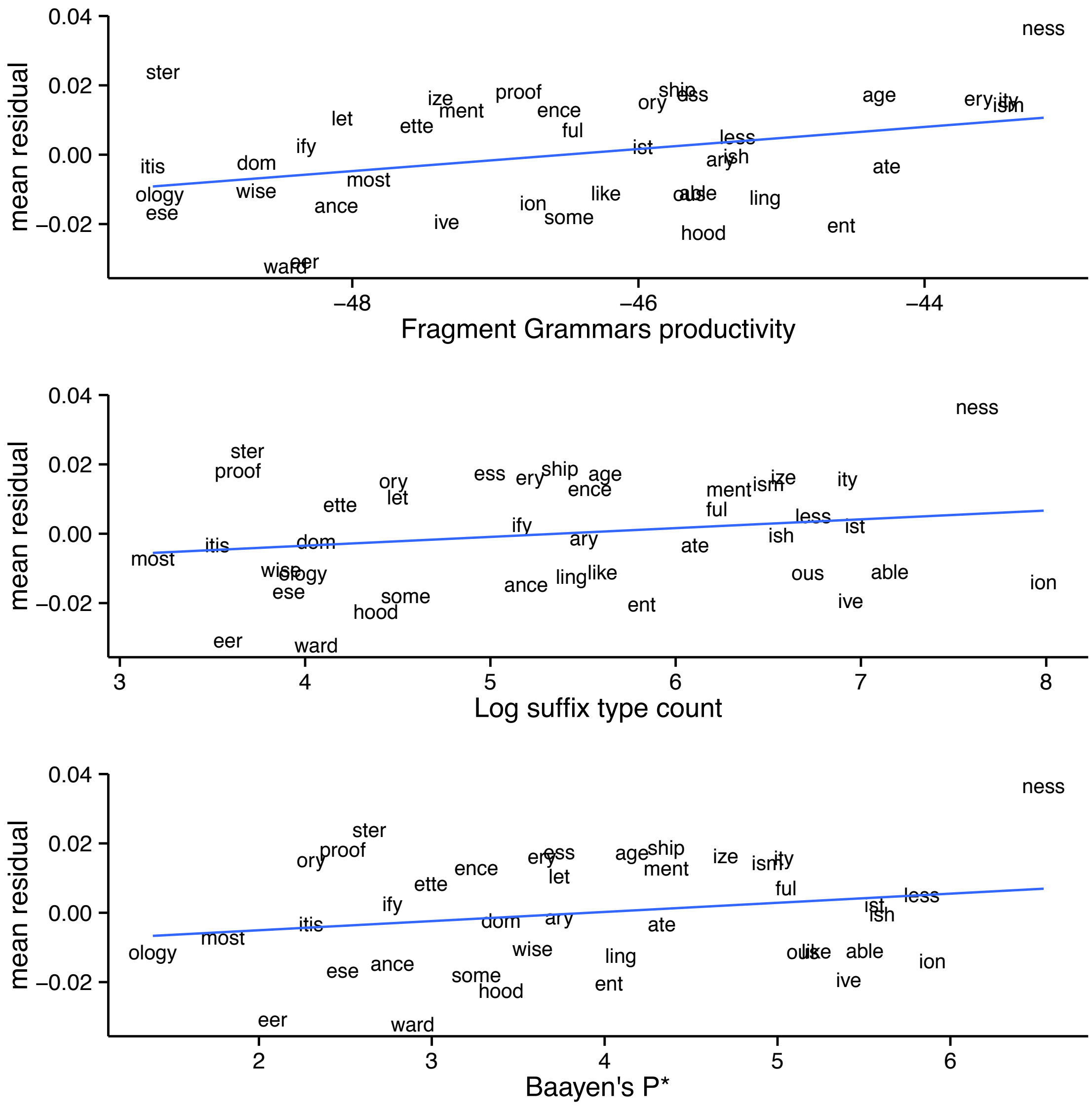

Figure 2 Each plot shows a different measure of productivity on the x-axis (Fragment Grammars productivity score, log type frequency of suffix, and Baayen's $\mathcal{P}^*$) and the mean residuals from the model by suffix on the y-axis. The upward trend from left to right shows an effect of productivity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

We also find the same pattern of significant results in an analysis of the data from Experiment 1 when we focused on just the full-suffix frames tested in Experiment 2. Using the same likelihood ratio test as before and only the full-suffix frames from the Experiment 1 data, we found that the model is significantly improved by adding a predictor from one of the three productivity models: for Fragment Grammars χ^2 (1) = 8.5, p < .01; for Baayen's $\mathcal{P}^*$: χ^2 (1) = 7.7, p < .01; for log type frequency of the suffix: χ^2 (1) = 9.2, p < .01). Thus, the generalization replicates in two data sets with different participants.

4. Discussion

Tversky and Kahneman (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 *accessibility* of linguistic representations during memory access. In this paper, we have proposed that the accessibility of existing linguistic structures is related to their productivity: the probability that the structure will be used to produce or comprehend a novel word. The connection between productivity and accessibility is simple: To comprehend a novel combination of word parts, like *tweet+able+ity,* language users must be able to independently access each word part from memory. Productivity is a measure of the probability with which word parts have independent representations independent of the words in which they appear.

Our experiments replicated Tversky and Kahneman's original findings for a much wider range of stimuli. As Tversky and Kahneman found, full suffixes produce higher frequency estimates than partial suffixes. More importantly, we have shown that *within* full suffixes, linguistic *productivity* explains differences in overestimation rates. Thus, overestimation rate provides a behavioral measure of productivity and join an existing body of literature on behavioral correlates of morphological decomposition.

Moreover, the results here present a linguistically motivated explanation for why Tversky and Kahneman find that partial suffixes are less accessible than full suffixes. Partial suffixes are, in effect, the ultimate low-productivity linguistic units: they never need to be accessed on their own. Therefore, people give strings with partial suffixes lower frequency estimates than the veridical frequency of these strings demands because of the accessibility bias. Similarly, low-productivity full suffixes are given lower frequency estimates than high-productivity suffixes for a similar reason: people use not the true real-world frequencies of these word endings but an accessibility-biased estimate.

Insofar as people's behavior on these linguistic tasks can be predicted by the need of accessing a particular structure as an unit, our findings fit well with general rational accounts of memory organization (e.g., Anderson and Milson, 1989; Anderson, 1990; Anderson and Schooler, 1991; McClelland and Chappell, 1998; Shiffrin and Steyvers, 1997; Hemmer and Steyvers, 2009; Steyvers and Hemmer, 2012) that suggest that *need probability* is a crucial factor in human memory and cognition more generally.

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. KM received support from National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a.

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.

Aronoff, M., & Schvaneveldt, R. (1978). Testing morphological productivity. *Annals of the New York Academy of Sciences*, *318*(1), 106–114.

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.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. Retrieved from http://CRAN.R-project.org/package=lme4

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

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

Bozic, M., Marslen-Wilson, W. D., Stamatakis, E. A., Davis, M. H., & Tyler, L. K. (2007). Differentiating morphology, form, and meaning: Neural correlates of morphological complexity. *Journal of Cognitive Neuroscience*, *19*(9), 1464–1475.

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.

Carroll, J. S. (1978). The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology*, *14*(1), 88–96. doi:10.1016/0022-1031(78)90062-8

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 648). Cambridge University Press New York.

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.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720. doi:10.1037/0003-066X.58.9.697

van Marle, J. (1990). Rule-creating creativity: analogy as a synchronic morphological process. *Contemporary Morphology*, *49*, 267.

Marslen-Wilson, W. D., & Tyler, L. K. (2007). Morphology, language and the brain: the decompositional substrate for language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 823–836. doi:10.1098/rstb.2007.2091

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.

O'Donnell, T. J. (2015). Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage. The MIT Press, Cambridge, Massachusetts and London, England.

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

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology*, *14*(3), 237–255.
the future: the prospective brain. *Nature Reviews Neuroscience, 8*(9), 657-661.

Shedler, J., & Manis, M. (1986). Can the availability heuristic explain vividness effects? *Journal of Personality and Social Psychology*, *51*(1), 26–36. doi:10.1037/0022-3514.51.1.26

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.