On hapax legomena and morphological productivity

Janet B. Pierrehumbert
Dept. of Engineering Science
University of Oxford
janet.pierrehumbert@oerc.ox.ac.uk

Ramon Granell
Dept. of Engineering Science
University of Oxford
ramon.granell@oerc.ox.ac.uk

Abstract

Quantifying and predicting morphological productivity is a long-standing challenge in corpus linguistics and psycholinguistics. The same challenge reappears in natural language processing in the context of handling words that were not seen in the training set (out-of-vocabulary, or OOV, words). Prior research showed that a good indicator of the productivity of a morpheme is the number of words involving it that occur exactly once (the *hapax legomena*). A technical connection was adduced between this result and Good-Turing smoothing, which assigns probability mass to unseen events on the basis of the simplifying assumption that word frequencies are stationary. In a large-scale study of 133 affixes in Wikipedia, we develop evidence that success in fact depends on tapping the frequency range in which the assumptions of Good-Turing are violated.

1 Introduction

The productivity of a morpheme is understood as the extent to which a language uses it actively in novel combinations. This vexed concept has multiple interpretations, of which two will concern us here. One views productivity as the cognitive propensity to create a new word involving a morpheme. The other infers productivity from the likelihood that new word types with a morpheme will be found when a corpus is expanded. The two can differ because the likelihood of finding a word depends not only on its creation, but also on the extent to which the word is learned and reused by others, and ultimately noted by an observer. One might suppose that a morpheme found in many different combinations would be more flexible in entering into novel ones, as in the rationale for Witten-Bell smoothing, (Jurafsky and Martin, 2000); if so, the type count of the morpheme would be a good index of productivity. However, the type count correlates poorly with human intuitions about productivity and with the number of OOV words found in test sets (Baayen and Lieber, 1991; Baayen and Renouf, 1996; Anshen and Aronoff, 1999). Working with corpora that are small by current standards, corpus linguists in the 1990s observed that the number of *hapax legomena* (or *hapaxes*) that contain a given morpheme is a much better predictor (Baayen and Lieber, 1991; Baayen and Renouf, 1996). This finding is argued to follow from assumptions about the cognitive system that make Good-Turing smoothing applicable, which we explain in the following section.

This paper systematically explores hapax counts as an indicator of productivity for a set of 133 morphemes that meet objective inclusion criteria for a much larger corpus than was used previously. This is the August 2013 download of Wikipedia that has 1.24 billion word tokens. We address several questions: Is the measure successful when exercised at a larger scale? Are the simplifying assumptions put forward to justify the measure valid? What does the behaviour of the measure tell us about the lexical system? We address these questions with numerical experiments. We define "pseudohapax" sets as sets of words in the full corpus that would be expected to occur exactly once in five nominal corpora having sizes used in classic studies. We explore how well the pseudohapax sets predict the distribution of morphemes amongst extremely rare words. We also downsample the corpus to create hapax sets from subcorpora matching the nominal corpus sizes. This approach allows us to separate the influence of several factors: sparse sampling, variation across morphological families in the shape of the rank-frequency distribution, and the actual frequencies of words that appear as hapaxes in corpora of classic size.

2 Hapax Legomena and smoothing

Hapax counts are advanced as an indicator of productivity in Baayen & Lieber (1991). The article describes as "large" the 18-million word Cobuild corpus (Renouf, 1987) on which the study was based. Hay and Baayen (2002) used the same corpus. The indicator has been widely used for more than 25 years (Baayen and Renouf, 1996; Chitashvili and Baayen, 1993; Plag et al., 1999; Popescu and Altmann, 2008; Kenny, 2014; Aronoff and Lindsay, 2014; Stump, 2017, in press). Several different measures can be defined from the hapax count. Of particular interest is the hapax percentage $P*$ (the percentage of all hapaxes that contain the morpheme). Using $P*$, the number of unseen word types with any given morpheme is estimated as proportional to its representation amongst the hapaxes, and this supports predictions about the distributions of morphemes in OOV sets much bigger than the hapax sets.

A justification for the hapax-based measures is proposed in Baayen & Renouf (1996) and Hay and Baayen (2002). They assume that the text meets the assumptions for Good-Turing smoothing: each word is produced with a constant probability, or put differently, the word frequencies are stationary and the text results from a Poisson process (Church and Gale, 1991). They also assume that the hapaxes are so rare that they are very likely to have been created on the spot. Integrating these assumptions, the idea is that the probability of creating a rare word form like *zeitgeist+y* or *post+Sumerian* is constant and based on the mental representation of the parts. The hapax set is by definition a subset of the word types with a morpheme, but nonetheless supports better predictions. The authors suggest the hapax measure works well because it eliminates complex words that are not decomposed during lexical access, and therefore do not contribute to productivity. Our numerical experiments were designed to include only words that are decomposable.

It is well-known that word frequencies fluctuate with the topic of discussion. Such deviations from a Poisson process provide the foundation for modern document retrieval algorithms (Sparck Jones, 1972; Church and Gale, 1995). The effects can be as large as orders of magnitude in frequency, and impact all parts of speech, although the impact on proper nouns is generally greatest (Church, 2000; Altmann et al., 2009). Different corpora can thus yield different $P*$ values for the same affix, providing the grist for post-hoc interpretation as in Plagg et al. (1999). In interpreting our results, we will also be concerned with the possibility of variability across speakers, which is of similar magnitude (Altmann et al., 2011).

3 Materials

In selecting our materials, one goal was to compare all morphemes that met objective inclusion criteria (as opposed to using subjective judgment to make a selection). The inclusion criteria were designed to identify morphemes that are reasonably familiar and that are reliably identifiable within complex words with a minimum of false positives. We considered words to be potentially decomposable into Prefix+Stem or Stem+Suffix if removing the affix yielded a stem that also occurs independently as a word of higher frequency. This criterion is needed to eliminate many spurious decompositions, such as *season* = *sea+son*, as well as words that are probably not decomposed into their parts in lexical processing (Hay, 2001). To select the target affixes, we began by considering all 184,499 words in Wikipedia that occur at least 100 times. Initial and final substrings of three or more letters were considered as potential affixes. The selected affixes occurred at least 50 times in the candidate list, and we also required that removing them reliably yield a valid stem. 68 prefixes and 65 suffixes met the criteria. The total set excludes many productive morphemes that coincidentally occur in many simplex words. It includes many true prefixes and suffixes, including combinations such as *+ingly, +ization, +ers*. Justification for treating these as units can be found in Stump (2017; in press). It also includes words that are used productively in compounding; the distinction between derivational morphology and compounding is a fuzzy one (Bauer, 2005). Detailed inclusion criteria and descriptive statistics, and a complete word list, are in the supplement (posted on the first author's web site).

Our outcome measure for productivity is the type frequencies for each morpheme family in the "far tail" of the distribution, defined as the set of words occurring 2 to 11 times. As is common in work on very large corpora, forms occurring only once are not considered because of problematic text normalization artifacts. The upper cutoff of 11 was selected to provide a large test set of words

	Prefix	Example	Types
MOST	non+	non-threat	11360
	anti+	anti-badware	4933
	sub+	subcritical	3520
	post+	post-vulgate	2854
LEAST	north+	northlands	428
	fore+	forebrain	410
	south+	southback	400
	second+	second-degree	365
	Suffix	**Example**	**Types**
MOST	+ers	adopters	6881
	+man	beckman	6639
	+based	rock-based	6169
	+like	garlic-like	5518
LEAST	+ful	needful	430
	+water	floodwater	399
	+american	austral-american	335
	+shire	dorsetshire	246

Table 1: Prefixes and suffixes having the most and least types in the far tail. Total types in the dataset.

with low frequencies (under 0.01 per million) that would be novel to many or most readers. The test set provides a much stronger mirror of underlying productivity than modest extensions of small corpora could. The far tail contains 129,714 complex word types that are relevant, in that they begin or end in one of the target affixes, and can be parsed into the affix plus a stem. Table 1 shows the most and least productive morphemes, as indicated by their counts in the far tail.

4 Frequency bands

We frame the calculations by considering nominal corpus sizes of 0.25%, 0.5%, 1.0%, 2.0%, 4.0% of the actual corpus size. Compared to classic corpus sizes, these range from rather small (3M words) to rather large (50M words). For each size, we take the "pseudohapaxes" to be words whose expected frequency in the nominal corpus would be 1.0, taking rounding into account. For example, for the 1.0% condition, the band is centered on words occurring 100 times. The 4.0% condition provides the largest possible pseudohapax set that has no overlap with the far tail (the test set).

The number of pseudohapaxes grows with the nominal corpus size. To evaluate the importance of sample size versus the absolute location in the frequency range, we also define down-bands, which have the same number of word types as a pseudohapax band but are simply shifted downwards in the frequency range by an integer multiple of the size of that band. Table 2 summarizes the banding scheme.

Size	PBand	PTypes	DB	HTypes
0.25%	[200,600]	5734	8	8196
0.5%	[100,300]	8463	5	11769
1.0%	[50, 150]	12709	2	16550
2.0%	[25,75]	19051	1	22886
4.0%	[12, 37]	29131	0	30650

Table 2: Banding scheme for five conditions, expressed as a percentage of the total corpus size. Frequency band for the pseudohapaxes (PBand), total number of pseudohapax types containing any of the prefixes or suffixes (PTypes), number of down-bands available before reaching the far tail (DB). Average number of hapaxes (HTypes).

For the 4.0% condition, there is no down-band, because the pseudohapax band falls just above the far tail. It is also important to look at the set of words with higher frequencies than the pseudohapaxes. In this "up-band" we include all words up to the most frequent; the size of the up-band is always within 15% of the size of the pseudohapax set. It is never possible to define more than one up-band from each set of pseudohapaxes.

A real hapax set corresponding to one of our nominal sizes would have only a sparse sample of the pseudohapaxes, but would also include words of higher and lower frequency. For each of the five pseudohapax bands, we simulate a real hapax set by taking a random sample of sentences in the corpus and collating the hapaxes. For each corpus size, 10 different subcorpora were created. If the hapax set happened to include words from the far tail, these were removed from the far tail for testing.

5 Evaluating predictions

We use ordinary least squares regression (OLS) to predict the logarithm of type count for each morpheme in the tail as a function of the logarithm of type count in a pseudohapax band, treating prefixes and suffixes separately. To ensure that the observations are robust, we use a hold-one-out method. Each prefix (or suffix) is held out and the remaining prefixes (or suffixes) are used to predict its value. We make the same calculation for all conditions, all down-bands and up-bands, and all hapax sets. Predicted R^2 values are adjusted

as described in Draper et al. (1998), yielding the measure $\bar{R}^2_{pred}$.

Table 3 summarizes the regression parameters. The slope is close to 1.0 for all fits (and closes in on 1.0 as the nominal corpus size increases), while the intercept varies. This means that the number of types in the far tail is approximately proportional to the number of types in the pseudohapax or hapax set, with the proportion decreasing as the nominal corpus size increases.

Figure 1 shows the results for $\bar{R}^2_{pred}$. For all conditions, the prediction from the pseudohapax band is better than the prediction from the up-band. Attempting predictions from words with frequencies over 600 (leftmost points in the figure) yields poor $\bar{R}^2_{pred}$ values of 0.4 and below. The pseudohapax bands for the 2.0%, and 4.0% conditions provide very good predictions with $\bar{R}^2_{pred} > 0.85$. This outcome is not chiefly due to the large size of these two pseudohapax sets. Predictions are nearly as good from the down-bands falling in the same frequency range. Figure 1 also shows results for the same calculation for the hapax sets; the reported $\bar{R}^2_{pred}$ averages over the results for the 10 subcorpora of each size. For the smallest corpora, the prediction from the hapax set is better than the prediction from the corresponding pseudohapax set. The hapax set is dominated by rare words, because lexical rank-frequency distributions are heavy-tailed. The median word frequency for the 0.25% case is 88 (or 0.07 per million). As we go towards larger corpora, the difference between the hapax value and the pseudohapax value dwindles.

6 Interpretation

No matter whether the sample is obtained as a hapax set or a pseudohapax set, success in predicting word types in the far tail depends on having a sample that is dominated by rare words. Psycholinguists view words with frequencies of 1 to 3 per million as low-frequency words, as in Carreiras et al. (2006), but a median of 0.07 per million was needed to achieve $\bar{R}^2_{pred} > 0.8$. Why did this outcome occur? Figure 2 sheds light on this question. It shows a frequency-rank distribution on a log-log scale for the 5 most productive, and the 5 least productive, suffixes as measured by the count of word types in the far tail. This is a rotation of a Zipfian rank-frequency distribution, with a separate sub-lexicon for each morpheme.

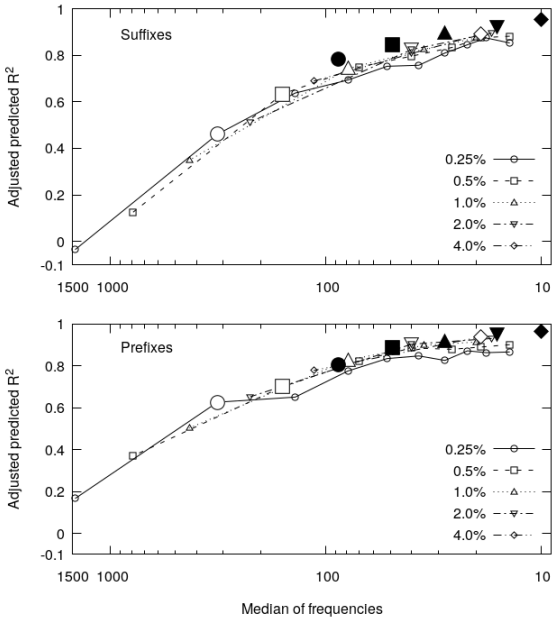

Figure 1: $\bar{R}^2_{pred}$ for using type counts of the affixes in the indicated band to predict type counts in the far tail. The pseudohapax set for each condition is indicated with an enlarged plotting character. The hapax set for each condition is indicated with a filled plotting character.

If a frequency spectrum obeyed a power law (as proposed by Zipf) it would appear as a straight line on a log-log plot. All curves are concave downwards, as typically observed (Baayen, 2001). There are marked differences in how the spectra roll off. Words with frequencies above 600 (0.5 per million) provide little information about productivity, and two of the most productive suffixes (*+like, +related*) still have not pulled out of the bottom group by 600. The slope around 100 is, however, very indicative of the slope around 10.

With a frequency of 88, the median hapax in the 0.25% case has a rank of 187,474 in the rank-frequency distribution for the entire Wikipedia vocabulary (not shown). This number can be interpreted in the light of results on adult vocabularies. Based on a large crowdsourcing experiment, Brysbaert et al. (2016) estimate that a 60-year-old at the 95th percentile of vocabulary knowledge knows 56,400 lemmas, or 95,880 words including inflected forms. Thus, it seems that unlikely that even such a person knows all of the hapaxes. Brysbaert et al. (2016) however omit proper names. So it is also relevant to consider "alphabetic words", which are words spelled with alphabetic characters regardless of their morphological status. Brys-

		Slope			Intercept		
	Size	Mean	Min	Max	Mean	Min	Max
Pseudohapaxes	0.25%	0.92	0.78 +based	1.05 over+	3.32	2.96 over+	3.70 +based
	0.5%	1.03	0.90, +based	1.15 wiki+	2.53	2.15 wiki+	2.92 +based
	1.0%	1.05	0.93 +based	1.15 side+	2.06	1.69 side+	2.47 +based
	2.0%	1.03	0.95 +like	1.10 second+	1.73	1.50 second+	1.99 +like
	4.0%	1.01	0.96 +based	1.06 home+	1.41	1.22 home+	1.60 +based
Hapaxes	0.25%	1.05	0.95 +like	1.16 self+	2.49	2.14 home+	2.78 non+
	0.5%	1.06	0.96 +like	1.15 side+	2.06	1.68 side+	2.42 +like
	1.0%	1.04	0.96 +like	1.11 news+	1.78	1.49 news+	2.11 +like
	2.0%	1.03	0.96 +based	1.11 non+	1.45	1.09 non+	1.76 +based
	4.0%	1.01	0.96 +based	1.06 head+	1.20	0.99 +ful	1.43 +based

Table 3: Summary of regression parameters. For minimum and maximum values, the indicated affix is the one that was held out.

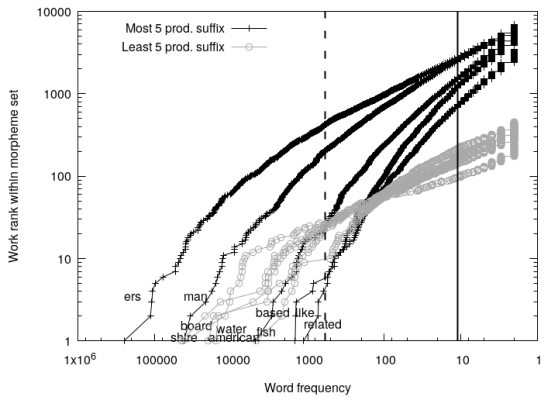

Figure 2: Frequency-rank distributions (on a log-log scale) for the most and least productive suffixes. Far tail to the right of the solid line at 11. 0.25% up-band to the left of the dashed line at 600. A comparable plot for prefixes is similar.

baert et al. (2016) apply the model fits in Gerlach & Altmann (2013) to estimate the number of distinct alphabetic words that a person has encountered, as a function of the total hours spent reading in their lifetime. For the Wikipedia editors, who had a median age of 25 in 2010 (Glott et al., 2010), reading 8 hours a day from age 5 yields a median estimated exposure to 146,000 alphabetic word types, which is still fewer than the median hapax rank. In short, the success of a hapax set as a predictor for words in the far tail depends on having words that are too rare to be known by everyone, and are therefore not constant in frequency across speakers.

We now consider the assumption that each word in the 0.25% hapax set was independently (and repeatedly) created with some probability. While this may be true for some words, it appears highly implausible for others. This frequency range includes many words that are not fully transparent and that recur many times within individual articles on specialized topics. Technical terms like *interaural* (audiology), *piquette* (oenology), *demand-side* (economics) are prototypical examples of words with non-stationary probabilities (Church and Gale, 1995; Curran and Osborne, 2002). For proper names, the suffix *+ville* is 17 times as productive as the suffix *+shire*. Given that Wikipedia asks all articles to be supported by secondary sources, few if any proper names would have been created on the spot.

We have seen that the hapaxes in a random sample of merely 3M words succeeded well in predicting the morphological profile in the tail of a corpus 400 times larger. The success seems to have occurred because the hapaxes provided a good slice of rare words that are not known to everyone, and that were not necessarily created on the spot. Pseudohapax sets that obtained a slice of similarly rare words worked just as well. Why are such rare words better indicators of productivity than more frequent words, even when these have been -filtered to be decomposable, as in this study? Possibly, rare words have a higher impact in ongoing learning of morphology because they are unexpected and salient. An alternative possibility brings in a social component. Different groups of editors in Wikipedia work on different topics. They may extend the morphological patterns that typify their field and distinguish it from other fields. In future research, we will evaluate such possibilities.

References

Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS one*, 4(11):e7678.

Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. 2011. Niche as a determinant of word fate in online groups. *PloS one*, 6(5):e19009.

Frank Anshen and Mark Aronoff. 1999). Using dictionaries to study the mental lexicon. *Brain and Language*, 68:16– 26.

Mark Aronoff and Mark Lindsay. 2014. Productivity, blocking and lexicalization. *The Oxford handbook of derivational morphology*, pages 67–83.

Harald Baayen and Rochelle Lieber. 1991. Productivity and english derivation: A corpus-based study. *Linguistics*, 29(5):801–844.

R Harald Baayen. 2001. *Word frequency distributions*, volume 18. Springer Science & Business Media.

R Harald Baayen and Antoinette Renouf. 1996. Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, pages 69–96.

Laurie Bauer. 2005. The borderline between derivation and compounding. In *Morphology and its demarcations: Selected papers from the 11th Morphology meeting, Vienna, February 2004*, volume 264, page 97. John Benjamins Publishing.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016. How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant?s age. *Frontiers in psychology*, 7.

Manuel Carreiras, Andrea Mechelli, and Cathy J Price. 2006. Effect of word and syllable frequency on activation during lexical decision and reading aloud. *Human Brain Mapping*, 27(12):963–972.

Revas J Chitashvili and R Harald Baayen. 1993. Word frequency distributions of texts and corpora as large number of rare event distributions. In *Quantitative text analysis*, pages 54–135. Wissenschaftlicher Verlag.

Kenneth W Church. 2000. Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p 2. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 180–186. Association for Computational Linguistics.

Kenneth W Church and William A Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech & Language*, 5(1):19–54.

Kenneth W Church and William A Gale. 1995. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.

James R Curran and Miles Osborne. 2002. A very very large corpus doesn't always yield reliable estimates. In *Proceedings of the 6th conference on Natural Language Learning-Volume 20*, pages 1–6. Association for Computational Linguistics.

Norman R Draper and Harry Smith. 1998. *Applied regression analysis*. Wiley series in probability and statistics: Texts and references section. Wiley, New York, NY.

Martin Gerlach and Eduardo G Altmann. 2013. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3(2):021006.

Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. 2010. Wikipedia survey ? overview of results. *United Nations University*.

Jennifer Hay. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics*, 39(6; ISSU 376):1041–1070.

Jennifer Hay and Harald Baayen. 2002. Parsing and productivity. In *Yearbook of Morphology 2001*, pages 203–235. Springer.

Daniel Jurafsky and James H Martin. 2000. *Speech and Language Processing*. Prentice-Hall, Upper Saddle River, NJ.

Dorothy Kenny. 2014. *Lexis and Creativity in Translation: A corpus-based approach*. Routledge.

Ingo Plag, Christiane Dalton-Puffer, and Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language & Linguistics*, 3(2):209–228.

Ioan-Iovitz Popescu and Gabriel Altmann. 2008. Hapax legomena and language typology. *Journal of Quantitative Linguistics*, 15(4):370–378.

Antoinette Renouf. 1987. Corpus development. In John M. Sinclair, editor, *Looking up: An account of the COBUILD Project in lexical computing*, pages 1–40. William Collins Sons and Co. Ltd. London, England.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Gregory Stump. 2017. Rule conflation in an inferential-realizational theory of morphotactics. *Acta Linguistica Academica*, 64(1):79?124.

Gregory Stump. in press. Some sources of apparent gaps in derivational paradigms. *Morphology*.