university of groningen

A quantitative approach to affix productivity

—

Estimating the reliability and validity of Twitter data

January 2019

AUTHOR: R. S. S. J. Buurke
COURSE: Corpus Linguistics (LTR024M05)
TEACHER: A.W. van Cranenburgh

1 Background

One of the many facets of language innovation is the formation of new words through morphological manipulation of existing words. Traditionally, studies into this so called *productivity* of particular morphemes have mostly been qualitative in nature, but with the rise of large corpora there have also been quantitative approaches to this phenomenon. The productivity of a morpheme can be seen as a property of a morpheme, which pertains to whether infinitely many new words can be formed by attaching the morpheme to existing words. Unproductive morphemes do not have this property, which is often the case for morphemes found in i.a. loan words, because the morphological licensing environment for the morpheme in language A is different than for language B. In other words, in French *-teur* may be productive, while it is not so in Dutch. Scholars do not always agree on whether or not the reality is simple, however, and many question whether a morpheme can be truly fully unproductive or productive. The distinction may be of a more gradual nature, i.e. the productivity of morphemes being relative to other morphemes. This is the assumption followed in this study, because to a certain degree it is impossible to capture a distinction that is clearly qualitative with data that is continuous and quantitative unless some arbitrary cut-off point is devised. There is little reason to believe that such a cut-off point will work across corpora, however, so the morphemes in this study will be strictly compared against each other, and no further generalizations are made pertaining to other morphemes.

A prime example of a quantitative approach to morphological productivity can be found in Baayen & Lieber (1991), in which a productivity measure was devised, which is termed *P* in this investigation.[1] It has been applied, sometimes in slightly different forms, in numerous subsequent studies, e.g. Baayen (1992) and Baayen & Renouf (1996). The original study applied *P* to a large number of English derivational morphemes, and the authors were able to delineate morphemes that are deemed productive on the basis of the literature from those that are unproductive. The measure correctly approximates the intuitively high productivity rate of suffixal *-ness* as opposed to the low productivity of i.a. *-ian*, but note that Van Marle (1992) criticizes the results to a considerable extent. One such point of friction is the assumption that hapaxes (words that occur exactly once in a corpus) serve as a valid indication of neologisms, because they could also be rare words. This is, of course, a fair observation (especially since it makes intuitive sense that some popular neologisms might be picked up and frequent in any case), but it need not mean that a frequency based approach is necessarily wrong. As long as the asymmetry between sufficiently low frequent words and more frequent words pertains to, or maybe more accurately correlates with, a distinction in innovativeness, a frequency based approach can still work. The crux of the matter is then to find an accurate and reliable quantification of such a distinction, which, given effective heuristics and pre-processing, is not *a priori* impossible. To a limited extent, this paper explores this possibility by testing the *P* measure with different frequency ranges for detecting innovative items.

Baayen and colleagues mostly used corpora from the 1990s, which were already sufficiently large for the relevant analyses, but they still had other drawbacks: a lack of balancing or of spoken language data. Especially the latter shortcoming, which admittedly remains a problem at present due to the enormous effort required for such databases, is possibly

[1] It should be noted that several productivity measures are developed by Baayen and colleagues over the years. This one, according to Baayen, should be used together with another measure in order to be truly reflective of a general notion of productivity. For reasons of time and space, only *P* is taken into scope.

partially solved by using a corpus of Twitter data. Smartphone use is pervasive throughout society nowadays, and – as a first world nation – definitely throughout Dutch society. In fact, population penetration of smartphones has reached over 90% in 2018 (Centraal Bureau voor de Statistiek, 2019). This readily serves the current purpose, as the language used in tweets is reflective of personal speech and interpersonal communication (although note that only data is available from public accounts, so highly personal conversations are presumably not represented), which is especially useful for finding neologisms, because presumably these show up more creatively and often in this type of communication than in more formal contexts. One further advantage of the Twitter data is its sheer size. The data used for this study, the Dutch unigram Twitter data collected and compiled by Bouma (2015), contains approximately 6.65M types between 2011 and 2015, which is orders of magnitude larger than databases from the 1990s. These advantages make for an interesting evaluation of productivity measure *P*.

Naturally, many advantages bring related disadvantages. In the case of Twitter data, there is one of direct importance for the analysis to be carried out. The problem is that the data is very noisy by the large number of spelling mistakes found in tweets. This is a problem not easily dealt with, because spelling mistakes can be quite divergent, and there is no straightforward way to filter them out of the data. One option is to only keep entries that match some dictionary source, but (1) dictionaries vary considerably in terms of size and accuracy, and (2) this would automatically delete neologisms as well. Solving this problem requires a separate study, so no attempt at fully solving this problem will be done here. Instead, it is estimated how detrimental this problem is to the particular productivity measure *P*. Note that spelling mistakes may very well be problematic for the analysis carried out in general if enough errors are sufficiently low frequent, which can be expected to be true given the fact that spelling mistakes can occur anywhere in a word. Moreover, especially since there is a risk of typographical errors on smartphones, the "alternative versions" of a certain word are virtually limitless. The analysis will show *how* problematic this interference is for the reliability of *P*.

The general question to be answered in this paper has already implicitly been put forward, i.e. how reliable *P* is under the conditions imposed by Twitter data. There is no literature available that directly provides insights as to what can be expected, so instead some *ad hoc* predictions are as follows. Recall that there is a conundrum in terms of hapax legomena and whether they actually pertain to neologisms. Given the voluminous and noisy nature of the Twitter data, however, there is little reason to assume this pattern to hold reliably. The prediction is that higher occurrence values are necessary to successfully, *ceteris paribus*, extract neologisms. It can be assumed that taking higher occurrence values can filter out a reasonable amount of the spelling mistakes in the dataset, although structurally pervasive errors will still be present (e.g. *afwesigheid* instead of *afwezigheid* might occur more than once as an effect of phonological-orthographic similarity). Solving these intricacies to that extent is beyond the scope of this paper, however. Instead, different parameters (both exact values as well as ranges) will be used to approximate what are reasonable settings for detecting neologisms in this dataset.

Lastly, the morphemes that will be investigated in terms of productivity are the following Dutch suffixes. They are split into two categories: presumed productive and presumed unproductive morphemes. The suffixes that are taken to be productive are *-je*, *-heid*, and *-baar*. The first morpheme *-je* is a noun-noun converting element, which simply converts a

noun into its diminutive form, e.g. *stoel* into *stoeltje*. It is presumably the most productive morpheme across the board, because " nearly everyone with some knowledge of the Dutch language has been struck by the frequency of its use of diminutives" (Shetter, 1959, p. 75). They fulfill an expressive role, and are therefore expected to occur often in the spoken language reflecting Twitter data. The suffix *-heid* turns adjectives into nouns, such as *ongelijk* into *ongelijkheid*. When used creatively it creates concepts related to the meaning of the adjective, e.g. words such as *aangeschotenheid*, which pertain to a conceptualization or level of being tipsy. Suffixal *-baar* forms adjectives from verbs, and bears the meaning that a certain action can be performed, e.g. *falsifieerbaar* indicates that something can be falsified. Creative uses of this particular suffix include cases such as *skatebaar* (the ability to skate over something) and *inhuurbaar* (the possibility of hiring something or someone). The presumably unproductive morphemes are *-teit* and *-teur*. They are presumed as such because they mostly occur in loan words from French and Latin, e.g. *amateur* and *kwaliteit*. Especially the loan words from French originated in a time that the influence from France was still much more significant, which is no longer the case, which applies especially to words ending in *-teur*. For the case of *-teit* it can be argued that it is still (marginally) productive, but again its heyday can be assumed to be over, as it has been analyzed as replacing the suffix *-té* in French loan words (e.g. *université* $\rightarrow$ *universiteit*; see the relevant entry in Van Der Sijs et al., 2009). The language changes induced by this suffix took place during Middle Dutch, and were no longer active by the time of Modern Dutch, so it can indeed be considered (relatively) unproductive nowadays.

2 Methods

2.1 Baayen's *P*

The calculation of Baayen's *P* is as in (1). It relates the number of hapaxes in which the suffix occurs to the total number of, at least in this investigation, types. More informally, it expresses the probability that a new word contains the particular suffix when N words have been sampled. In terms of implementation the algorithmic steps are straightforward: for each item in the sample, add one to the sum of times this suffix was encountered in a word, and divide that number over the amount of words sampled at that point.

(1) $P = \frac{n_1}{N}$
For which n_1 is the number of hapaxes and N is the sample size.

There are a few notions that need to be addressed here. First off, given the fractional nature of the equation, and its dependence on the sample size (which is especially large for Twitter data), the values of *P* will always be small. As noted in Section 1, however, this does not constitute a problem in terms of interpretation, because all values are interpreted relative to each other and not in absolute terms. Moreover, as noted in Section 1, n_1 does not pertain to hapaxes in this particular investigation. Instead it refers to specific frequency ranges of occurrences of types. In particular, the ranges 0–20, 00–30, and 0–50 will be used. These ranges were chosen after inspection of the data, which yielded the insight that the Twitter data is truly too voluminous (especially after aggregating all the months into data per year) to expect frequency counts for any type to be lower than 10. A comparison of the suffixal

prevalence in these ranges will yield insight into how broad the range must be to detect the most neologisms with the minimal amount of noise.

Naturally, for each larger range more noise is generally to be expected. In order to quantify how much, a "neologism ratio" (henceforth N-ratio) is set up. It is defined as the number of neologisms out of 15 randomly sampled items per group divided over the non-neologisms. For example, if for the 0–20 range x number of types with a particular suffix are extracted, and $x > 15$, then it is manually counted how many clear neologisms (as opposed to noise) are found in that particular set. If $x < 15$, then it is simply counted and divided over all available words. The following cases are taken to be non-neologisms: (1) typographical errors, (2) spelling errors, (3) cases in which the suffix is not used as an active suffix, e.g. in the loan word *barbaar*, and (4) derivations in which the penultimate step is intuitively and clearly a non-neologism itself, e.g. *superbetrouwbaar* is taken as a non-neologism, as *betrouwbaar* is not a neologism. It is of course still a neologism, but not one relevant for the productivity of *-baar*.

2.2 Dataset and sampling

As noted above, the relevant Twitter database is that of Bouma (2015). The data per month for each year are freely available from the website. As there is no obvious reason to assume that the measure P will be more reliable for any of these closely related years the data from one year is used for the analyses: 2011. The data come in frequency counts per months, which were aggregated to form a single dataset consisting of all counts for that year.

The total 2011 corpus after aggregation consisted of approximately 706 thousand types, which were drawn from over 8.7M tokens. In order to reduce the computational load, as no computer cluster was available, and also to check whether there is a clear effect of sample size in general, it is worth testing whether the same measure P is found across different sample sizes. This is tested for randomly drawn samples of size 5k, 50k, and 500k for productive *-heid* and unproductive *-teit*. The results for each drawn sample are reported in Table 1.

Table 1
Values of P for different samples and sample sizes for the suffixes -heid and -teit (with a frequency range of 0–50.)

Suffix (sample size)	***P* after N sampled types**				
	100	5000	10000	50000	500000
-heid (5k)	0.01000	0.00060	-	-	-
-heid (50k)	0.01000	0.00060	0.00060	0.00036	-
-heid (500k)	0.01000	0.00020	0.00040	0.00036	0.00047
-teit (5k)	0.01000	0.00020	-	-	-
-teit (50k)	0.01000	0.00020	0.00040	0.00012	-
-teit (500k)	0.01000	0.00040	0.00020	0.00012	0.00008

It is clear that a random sample of 5k yields noticeably different results than that of 500k, especially when looking at the 500k sample of *-teit*. After 5000 sampled types the value of P is twice as large for the 500k sample as for the 5k one, which is counterintuitive and also rectified by the processing of 10000 types. A 50k sample size is clearly more attractive to use

than a 500k one in terms of required computation time, and it should be noted that the values seem to converge after 50000 types have been processed. However, it also is clear that the productivity distinction is best shown when a sample size of 500k is used, because the value for presumed productive *-heid* is then 0.00047 and that of unproductive *-teit* is 0.00008. Note also that the value of *P* rises again for *-heid* after 500000 types have been processed. This is an interesting finding, because as N gets infinitely large the value of P should approach 0. In other words: *P* is not necessarily monotonically decreasing in all cases, so processing more types is still a worthwhile investment. In conclusion, then, it must be concluded that it is the best option to simply use the full corpus of 706k types after all. 50k-sampling is already effective, but using more data is clearly significantly more informative.

3 Results

Table 2
Examples of neologisms for -je, -heid, and -baar.

-je	-heid	-baar
priegelwerkje	prikkelbaarheid	kotsbaar
zussendagje	verhevenheid	msnbaar
bloedprikje	gammelheid	bladerbaar
linkerknopje	onrechtmatigheid	zitbaar
priveetje	happyheid	coachbaar

Table 3
Values of P and N-ratios for each suffix by occurrence frequency of types.

Suffix	**Frequency range**					
	0–20		0–30		0–50	
	P	N-ratio	*P*	N-ratio	*P*	N-ratio
-je	0.00010	7/15	0.00120	7/15	0.00531	9/15
-heid	0.00002	3/13	0.00011	4/15	0.00045	10/15
-baar	<0.00000	1/3	0.00004	8/15	0.00015	10/15
-teit	<0.00000	0/1	0.00002	0/11	0.00009	3/15
-teur	<0.00000	-	0.00001	0/8	0.00005	0/15

Examples of neologisms extract from the Twitter data are reported in Table 2. Note that in the case of *-je* it may sometimes be questioned to what degree the neologism is to be considered truly innovative, as it does not seem to add much depth to the meaning to simply form diminutives. Recall however that these forms are mostly expressive, and therefore do add meaning, but on a pragmatic level. The neologisms of *-heid* and *-baar* are more clearly used creatively, and indeed form new concepts from adjectives and verbs respectively.

The results of the analysis are reported in Table 3. In terms of ranking productivity, the outcome is mostly as expected: *-je* is by far the most productive in terms of the *P* measure, followed by *-heid* and *-baar*, which are again more productive than *-teit* and *-teur*. The *P*

values of *-je* are between 5 and 11.8 times larger than next highest productive morpheme *-heid*, and at least between 10 and 106 times more productive than the least productive *-teur*. According to *P* ranking the suffixes *-heid* and *-baar* are in the middle in terms of productivity. They are respectively between 2 and 9, and between 1 and 3, times as productive as least productive *-teur*. Again as expected, *-teit* is still in most cases twice as productive as *-teur*. Given that it is indeed possible to apply *-teit* creatively, it is interesting to see that it happens relatively little compared to *-heid* and *-baar*. What this finding entails for the concept of productivity in general is discussed in Section 4.

In terms of N-ratios, an interesting pattern in terms of frequency range emerges. As can be expected, they are consistently decent for *-je*, which corresponds with its clear edge in terms of productivity. For *-heid* and *-baar* it is clear that the frequency range 0–20 is not reliable for extracting neologisms, as there are both few words within this frequency ranges, and even less neologisms. For the unproductive *-teit* and *-teur* it is even the case that no neologism is found. In fact, there are almost no occurrences of these suffixes when sampling within this frequency range. While the N-ratio is once again reasonable for *-baar* with a frequency range of 0–30, the same cannot be said for *-heid* with only 4/15 drawn types being clear neologisms. The highest chance of correctly extracting neologisms from the data, however, is to be found in the frequency range of 0–50. This holds across all productive suffixes, as well as *-teit*.

4 Discussion

In this investigation Baayen's *P* measure for morphological productivity has been applied to Dutch Twitter data from 2011. The data were not pre-processed due to constraints of time and space, which, while it is necessary to do so for higher accuracy rates, is nonetheless interesting as it provides a relatively direct comparison to the original studies in this area, especially since the databases used prior were also not pre-processed. Manual approximation of noise ratios, or more specifically non-neologism (the aforementioned N-ratios), were calculated to chart the influence of particular types of noise in Twitter data, e.g. typographical errors and spelling mistakes.

Interestingly, and encouragingly, from the results it becomes clear that this particular measure for morphological productivity (recall that Baayen and colleagues have developed others as well) seems to produce intuitive results. The expectation was that the Dutch diminutive-forming suffix *-je* would be the most productive, and that prediction was indeed borne out. Similarly, the *P* values for *-heid* and *-baar* indicate a level of productivity, but the case of *-baar* is interesting. Note that in absolute terms (for the preferred 0–50 frequency range at least) the difference between *-baar* and *-teit* is small. Even in relative terms *-baar* is only about 1.67 times as productive as *-teit*. There are two interpretations that can readily explain why this is the case. The first possibility is that there is simply much noise in the *-heid* data, which indeed seems to be the case given the low N-ratio of 3/15. The prevailing error type, from a quick glance at the data, seems to be typographical errors, e.g. *oveheid* instead of *overheid* and *tilheid* instead of presumably *stilheid*. It is not immediately clear why these errors are so often found for this particular suffix, and ostensibly less so for the other suffixes. Another option is to re-evaluate whether *-teit* is in fact as unproductive as was claimed in the Introduction. Recall that the assumption was that its heyday in terms of productivity was over,

as the suffix entered the Dutch language already in the Middle Dutch period. However, the suffix is indeed still productive to a certain degree, e.g. in rare cases of morphological analogy such as *stommiteit* and *flauwiteit*. It is clearly less productive than the presumed productive suffixes, but the fact that this subtlety seems to be picked up by the investigation is interesting in and of itself. It corroborates the conclusion that the *P* measure works relatively well for comparing the suffixes of this investigation, at least in relative terms.

One obvious shortcoming of the application of *P* as it has been done in this paper is that it is difficult to replicate reliably. The implementation of *P* is straightforward, but recall that for this particular study the formula for *P* is in fact actually radically different. For the original version of *P* as proposed by Baayen and colleagues, the measure depended on so called hapax legomena. These words that occur *exactly and no more than* once were hypothesized to correlate strongly with the occurrence of neologisms, but this concept had been criticized on multiple accounts by others and is simply not applied in this study. Low frequency ranges were used as an alternative correlate with neologism, which for the 0–50 worked out relatively well, but frequency ranges are highly dependent on the dataset. For Twitter data it can be expected that even greater ranges than 0–50 might still yield many extracted neologisms, but for a more balanced or carefully compiled corpus it can definitely be the case that a range of 0–50 is already too great and includes too many different lexical types. A task for future inquiry can be to estimate what the ideal frequency ranges are for specific corpora based on their properties.

It should also be noted that the N-ratios were at best 10/15. This result is not unsurprising given that the metric depends on ranges instead of particular values, but it indicates an inherent problem with the correlative assumption between neologisms and low frequency items. To be more precise: low frequency items (especially in noisy Twitter data) include words of a different nature, such as typographical errors, spelling mistakes, and simply rare words as well. Whether or not a singular value is used or not, there is little reason to assume that this problem does not persist. The original concept is therefore useful, but inherently flawed, and can never be expected to attain near-perfect N-ratios, unless the data is pre-processed extensively to filter out the other types of low frequency items. But if such significant pre-processing is required, one may also wonder what the actual contribution is of *P* in the process, as the pre-processing is clearly the more crucial part of the computation. In any case, it must be concluded that hapax legomena are not *the* crucial ingredient for neologism detection, and other concepts correlate with neologisms as well.

One final point to return to is the concept of morphological productivity in general. Recall that scholars do not always agree on whether there is a two-way distinction in the productivity of a morpheme or whether there is a gradual continuum. It was not the original intention of this investigation to directly add to this debate, but the results of one particular suffix instigate a short discussion: those of *-teit*. As noted earlier, this particular suffix has a unique status, because it is still theoretically productive, but the actual Dutch population does not seem to use it as such anymore. This raises the question of whether the theoretical possibility of producing new words with a particular morpheme is indeed what should be considered *productive*. Surely, productivity pertains to actual language use as opposed to a property that is not being exploited. Put another way: the concept of morphological productivity can be interpreted in different manners. The measure *P* clearly estimates the level of productivity in terms of actual language use, but if one desires to determine whether the property of productivity is to be assigned to a particular morpheme another method must be

put in place. In all likelihood this can only be truly achieved by qualitative methods, but this remains an open question to be answered by future studies.

5 References

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: A corpus-based study. *Linguistics, 29*(5), 801-844.

Baayen, H. (1992). Quantitative aspects of morphological productivity. In *Yearbook of morphology 1991* (pp. 109-149). Springer, Dordrecht.

Baayen, H., & Renouf, A. (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language, 72*(1), 69-96.

Bouma, G. (2015). N-gram Frequencies for Dutch Twitter Data. *Computational Linguistics in the Netherlands*, Journal 5, 25-36.

Centraal Bureau voor de Statistiek. (2019). Internet; toegang, gebruik en faciliteiten [Dataset]. URL: https://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=83429NED&D1=0,2-5&D2=0,3-6&D3=0&D4=a&HDR=T&STB=G1,G2,G3&VW=T

Lüdeling, A., Evert, S., & Heid, U. (2000). On Measuring Morphological Productivity.

Shetter, W. Z. (1959). The dutch diminutive. *The Journal of English and Germanic Philology, 58*(1), 75-90.

Van Der Sijs, N., Debrabandere, F., Philippa, M., & Quack, A. (2009). *Etymologisch woordenboek van het Nederlands*, online edition. Amsterdam: Amsterdam University Press.

Van Marle, J. (1992). The relationship between morphological productivity and frequency: a comment on Baayen's performance-oriented conception of morphological productivity. In *Yearbook of Morphology 1991* (pp. 151-163). Springer, Dordrecht.