Productivity, vocabulary size, and new words. A reply to Säily (2016). Draft version.
To appear in *Corpus Linguistics and Linguistic Theory*.

Kristian Berg, Universität Oldenburg
kristian.berg@uni-oldenburg.de

1. Introduction

Säily (2016) offers a very interesting way of gauging the productivity of a word formation pattern when the corresponding token counts vary. However, I disagree with the operationalization of productivity that she uses. In this short paper, I will a) show why Säily's method does not work for productivity changes, and I will b) sketch an alternative method that avoids these pitfalls. As a data set I will use the large diachronic corpus of German Deutsches Textarchiv (www.deutschestextarchiv.de, henceforth DTA); I will focus on two word formation patterns, *-isch* and *-nis*.

2. Types and tokens

It has long been known that the relationship between types and tokens is not linear (cf. e.g. Baayen 1992: 113); that means that type counts of subcorpora of different sizes cannot be compared. Säily's idea to solve this problem is simple and compelling: We randomly reshuffle the corpus a lot of times (1,000,000 times in Säily's paper), leaving the individual texts intact. For each iteration, we observe the type counts with every increase in the token count. From this large amount of re-sampled corpora, we can then determine which type/token ratios are normal, and which are not: For any given token count, we know how many types to expect on average, and which type counts deviate so much that the deviation is significant. As Säily points out, the method is essentially visual. We can plot the type/token ratios for the subcorpora we are interested in against the probability distribution.

With this method, we can divide the initial corpus according to all kinds of (socio-)linguistic variables; Säily investigates gender, social rank, and time span. However, Säily is not interested in type counts per se; she uses it as a measure of productivity, and this is where I object. Using type counts as an indicator of productivity is not unusual; the size of the vocabulary for a given word formation pattern, i.e., the distinct types, is what Baayen (2009: 901f.) calls the 'realized productivity'. But realized productivity is just one of three measures suggested by Baayen (1989, 1992, 1993). The other two measures involve counts of hapax legomena. Now it is true that these measures do not work well in small corpora like the ones Säily uses (cf. p. 147). But as a consequence, Säily is left with just one of Baayen's trio of measures — arguably the one that is least suited for diachronic investigations (it is important to keep in mind that Baayen's measures were initially developed for *synchronic* investigations of productivity).

Why is realized productivity not suited for diachronic investigations? Because with this measure, we can only determine whether a pattern is productive at all, not whether the level of productivity changes over time. Increases or decreases in productivity over time cannot be captured. To see why this is indeed the case, we need to step back a little. Word formation patterns can carry a lot of baggage: Old words that were once new but have since become commonplace are as much part of the vocabulary as are productively coined new

words. Word formation patterns thus tend to sediment types over time.[1] Baayen (2009: 901f.) puts it, realized productivity measures the "past achievement" of a word formation pattern — and we cannot distinguish between past and present achievements based on this measure alone. Judging from a list of types like *fraternity*, *ductility*, *security*, and *obscurity* (examples from Säily 2016:137), it is impossible to say which are institutionalized or lexicalized, and which are new. Imagine two vocabularies of equal size, with one containing only old words with *-ity*, and one only new words. The second one should be more productive than the first one. Yet with realized productivity, we cannot tell them apart. That is precisely what hapax-based measures set out to do: They take very rare words (in a large enough corpus) as an approximation to new formations.

We can state this problem differently: The dispersion of types (in the sense of Baayen 2001) over time is not equal; later time spans are favoured. Lexicalized words that predate the corpus have a higher probability of being attested throughout the corpus, while words that were coined at a later stage will of course not appear earlier in the corpus. This holds for patterns that are constantly highly productive throughout time, patterns that rise or fall in productivity, and also patterns that are marginally productive. Wholly unproductive patterns are the only exception to this: Their vocabularies will not grow. Here, the vocabulary size provides a hint. But unless we are interested in a dichotomic distinction between productive and unproductive, the measure is flawed because all productive patterns are biased towards later stages, and we cannot, for example, distinguish constant levels of productivity from rising levels.[2]

3. A case study: German *-isch* and *-nis*

So far, I have argued theoretically. Let me illustrate my point with data from a large diachronic corpus of German, the DTA corpus. I used Säily's method to compare the vocabulary size of the adjective suffix *-isch* (e.g. *alkoholisch* 'alcoholic') for the 19th century decades. I plot the number of distinct types (the vocabulary of the *-isch* pattern) against the number of running words.[3] Instead of 1,000,000 permutations, I used 100,000 (if anything, the results should be getting more significant if we reshuffle the corpus 900,000 times more). The result is the typical "banana-shaped plot" (Säily and Suomela 2017), together with dots that indicate the number of types and running words for *-isch* for the respective decades[4].

[1] Of course, words can also become obsolete, as one anonymous reviewer points out. On the whole, however, this erosion seems to be less important than the sedimentation of words (cf. e.g. Klein 2013 for data from German).

[2] As one anonymous reviewer rightly points out, the usual caveat applies: The language that a speech community uses is never fully represented in a corpus. The related question of how large a diachronic corpus has to be to yield reliable results is very interesting, but ultimately beyond the scope of this paper.

[3] We could also plot the number of distinct types against the number of tokens with *-isch*. However, Säily (2016: 148) points out that the results for tokens (instead of running words) are "similar […] but less significant": I therefore only used the running words measure.

[4] The Python scripts for the Monte Carlo simulations for this and the other plots in this paper can be found at github.com/kristian-berg/CLLT.

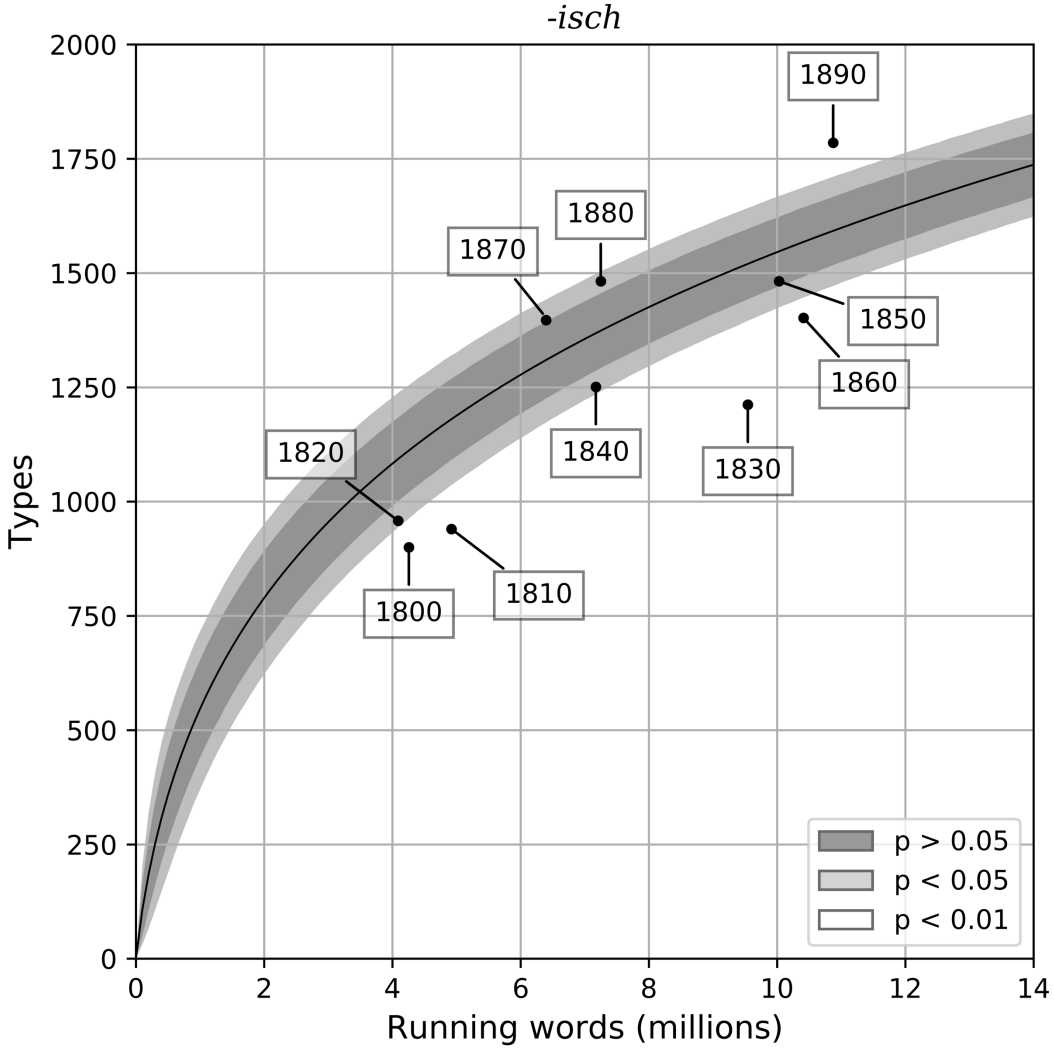

Figure 1: Säily plot for -isch. The line shows the mean vocabulary size for -isch, the shaded areas indicate the bounds ($p < 0.05$, $p < 0.01$) for 100.000 type accumulation curves, with subcorpora based on decades plotted on the curves. Data base: Deutsches Textarchiv.

We see that the token counts (in running words) of the decades vary substantially: For the 1820-29 decade, there are only about 4 million running words, for the 1890-99 decade, there are almost 11 million. This is the very reason we cannot simply compare the type counts. More importantly, we find that the distribution of decades is as expected: The six earliest decades (1800-1860) contain less types than the average of 100,000 randomly reshuffled corpora. Säily's method leads us to believe that this means the productivity of the word formation pattern has increased in the 19th century. But the data simply do not warrant such a statement. All we can deduce is that the pattern is not unproductive. As a matter of fact, when we use new words to measure productivity, we see that the productivity of this pattern is fairly constant throughout the 19th century (see figure 4 below).

We can show that this is an artefact of the method by looking at a second suffix, *-nis* (e.g. *Geständnis* 'the act of confessing'). *-nis* is a noun suffix that is generally considered to be

marginally productive at best. If we apply Säily's method to this word formation pattern, it yields essentially the same result.

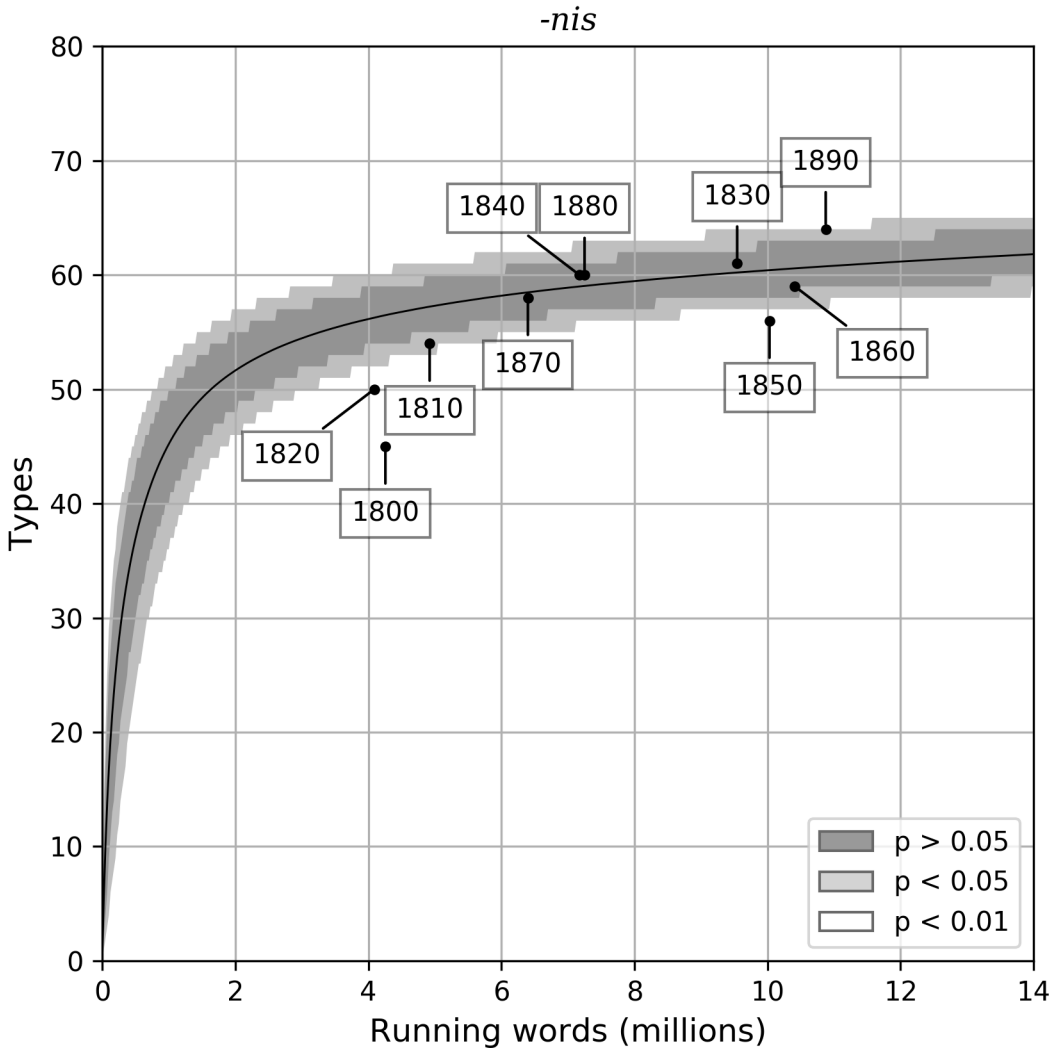

Figure 2: Säily plot for -nis. The line shows the mean vocabulary size for -nis, the shaded areas indicate the bounds ($p < 0.05$, $p < 0.01$) for 100.000 type accumulation curves, with subcorpora based on decades plotted on the curves. Data base: Deutsches Textarchiv.

Again, earlier decades contain significantly less types than the average reshuffled corpus, and later decades contain more types (with some exceptions). However, this pattern is only marginally productive, as I will show below. Yet applying Säily's method would lead us to state an "increase in productivity" (Säily 2016:136) of *-nis* over time.

To determine the relation between expected and observed values more systematically, figure 3 plots the difference between the mean re-samples values and the actual values for both patterns. Note that the numbers of running words in this figure are cumulative (for example, the decade 1860-69 is plotted at 50 million words because that is the total word count between 1800 and 1869).

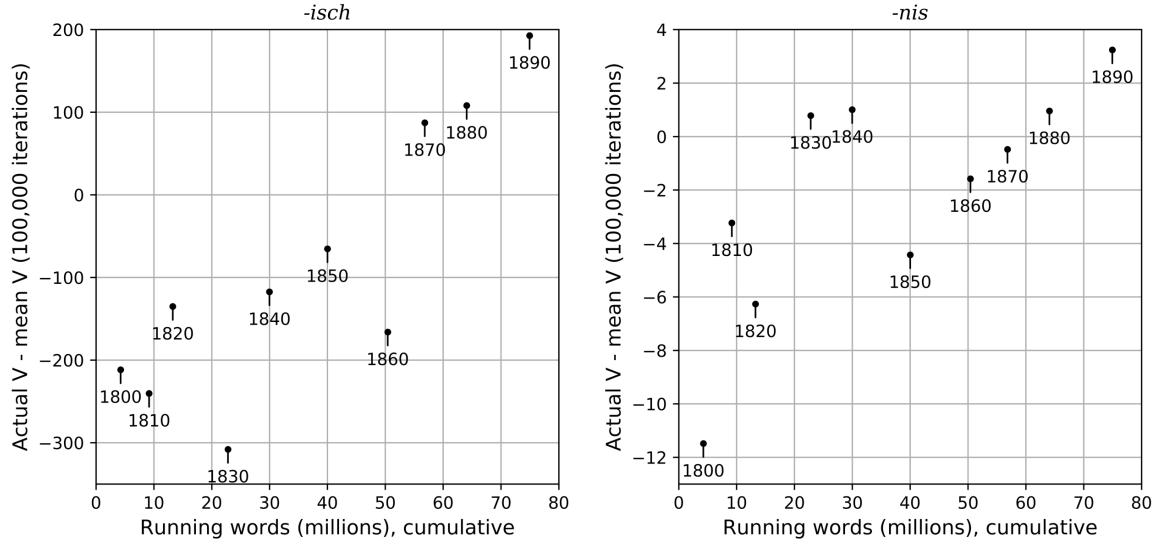

Figure 3: The difference between actual vocabulary size and expected vocabulary size (as computed by 100,000 random re-samplings of the corpus), plotted over the cumulative amount of running words for -isch (left panel) and -nis (right panel). Data base: Deutsches Textarchiv (DTA).

With only a few exceptions, earlier decades have a smaller vocabulary size than expected, and later time spans have a larger vocabulary size than expected. The difference between expected and observed values and the amount of running words are closely correlated (Spearman's rho for *-isch*: ρ=0.83, p = 0.0029; Spearman's rho for *-nis*: ρ=0.71, p = 0.0216).

This shows that for productive and unproductive patterns alike, distributions like those in figure 1 and figure 2 arise naturally: Some word formation products tend to fossilize over time (to use yet another metaphor), and that makes vocabulary size unsuitable as an indicator for morphological productivity. A growing vocabulary does not indicate an increase in productivity; a steadily productive word formation pattern leads to a steadily growing set of words. In a way, Säily falls prey to her own conceptualization of productivity as vocabulary.

4. New words

What is the alternative? The notion of productivity is closely linked to new words (cf. e.g. Baayen 1993: 183), and diachronic corpora allow us to extract information about the newness of words — so I suggest we use new words to determine diachronic productivity (for the following cf. Berg subm.). More precisely, we determine what proportion of all types of a pattern in any given decade are new types (P_{neo}). This is a very straightforward, direct measure, it is easy to operationalize and it is directly interpretable.

There are two methodological obstacles, however. The first one is the very reason for Säily's approach: Determining new words (with a given word formation pattern) per decade is reasonably easy, but the DTA corpus contains varying token counts for the decades, so we cannot simply normalize. Yet instead of discarding normalization altogether, I suggest we simply use a different method.

To this end, we determine a cut-off for the decades, the lowest common denominator so to speak. In the case of the DTA corpus, this size is around 4,000,000 running words for each

decade between 1700-1890. Then we use a Monte Carlo simulation: We randomly re-sample 4,000,000 running words from the corpus many times (similar to Säily's original method), count new words in each sample, and then determine the mean and the bounds beyond which only 5% and 1% of the data fall. This way, we make sure that no data in a given decade are neglected.

The second problem is that all word formation patterns tend to be more productive at earlier stages in the corpus (cf. e.g. Cowie and Dalton-Puffer 2002: 429; Kempf 2016: 116). The explanation is simple: All words that predate the corpus "must initially register" (Cowie and Dalton-Puffer 2002: 429). The higher levels of new words that we observe are (for the most part) not new at all, but older words that make their first appearance in the corpus. The levels can thus not be taken at face value. To overcome this problem, I suggest we use an earlier corpus as an indicator — in this case the Early High German corpus Bonner Frühneuhochdeutschkorpus[5] — and determine for each lexeme in this earlier corpus its first occurrence in the later DTA corpus. From this we can calculate the point in time when the probability of an old word from the Early High German corpus first occurring in the DTA corpus is so small it is negligible. This point in time, it turns out, is around 1700 (cf. Berg subm.). Accordingly, I only use DTA data after 1700. This way, the chance to wrongly encounter old words as new is rather low (<5%).

As introduced above, the suggested measure P_{neo} relates the mean number of new types per time-span (in our case, decade) to the mean number of all types. For example, there are on average 734 *-isch* types (old and new) in 1800-09, and 94 of them are (again, on average) new. That means we arrive at a productivity level $P_{neo} = \frac{94}{734} = 0.128$.

This measure is preferable because it is directly interpretable as the reproduction rate of the word formation pattern: In that decade, almost 13% of *-isch* types were new words. What is more, we can directly compare the P_{neo} values for different suffixes and determine which one is more productive (Säily's method only allows a direct comparison for aspects of one pattern). Figure 4 is a plot of the P_{neo} values for *-isch* and *-nis* over the course of the 18th and 19th century.

[5] This corpus contains 40 Early High German texts from the 14th to the 17th century with a total of around 600,000 tokens (http://www.korpora.org/fnhd/); cf. Lenders and Wegera (1982).

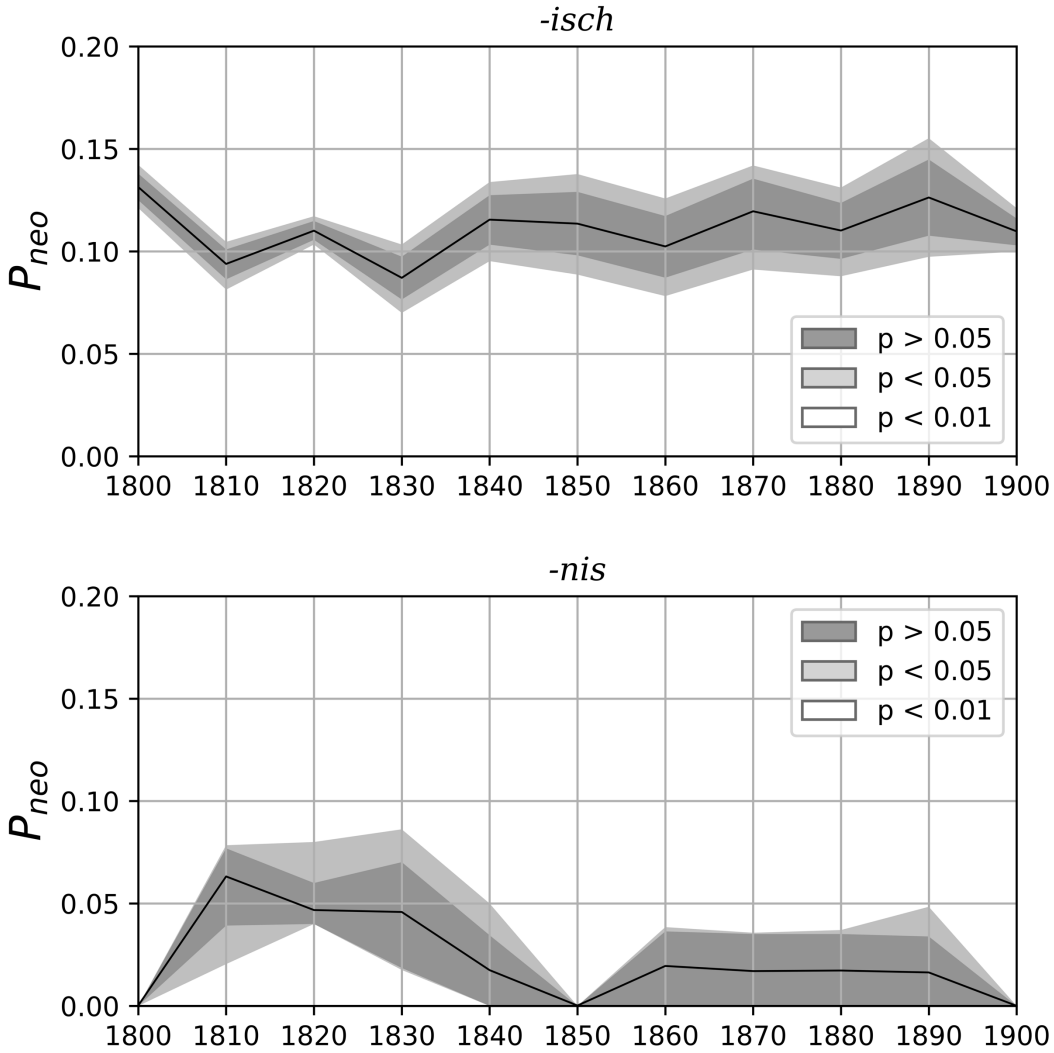

Figure 4: Mean P_{neo} values for -isch and -nis with with bounds ($p < 0.05$, $p < 0.01$) for 100.000 randomly resampled versions (with 4 million token per decade) of the corpus Deutsches Textarchiv (DTA).

According to this measure based on new types, *-isch* is constantly productive at a rather high level over the course of two centuries. This constant productivity leads to a constantly growing vocabulary. *-nis*, on the other hand, is considerably less productive, with P_{neo} values of zero in many decades.[6]

There are at least two questions regarding figure 4: Are the short-term fluctuations significant? And are there also patterns with a long-term change (rise or fall) in P_{neo} values, or is the measure always more or less static?[7]

As for the first question, consider the decline of P_{neo} for *-isch* between 1820 (0.11) and 1830 (0.087). Is this decline significant or random? It is significant because we can witness it in 99.6% of the 100,000 permutations. There are only 364 versions of the corpus where the

[6] Measures based on new types are also superior to hapax-based measures such as Baayen's $\mathcal{P}$, as I have argued elsewhere (Berg subm.). It is thus no remedy for Säily's approach to use hapax counts instead of vocabulary size (Säily 2016: 132).

[7] I would like to thank an anonymous reviewer for raising these questions.

levels do not drop. The decline in new words with *-isch* between 1820 and 1830, we can conclude, is a rather stable feature in the DTA corpus. This is of course a feature that is not readily observable from graphs like figure 4; it is certainly worthwhile to think about a way to visualize this information.[8]

The second question is whether the P_{neo} measure can also pick up gains in productivity. To show that this is indeed the case, compare figure 4 with the P_{neo} values for the noun suffix *-tum* (e.g. *Arbeitertum*, 'the collection of all workers [i.e., the working class]') in figure 5.

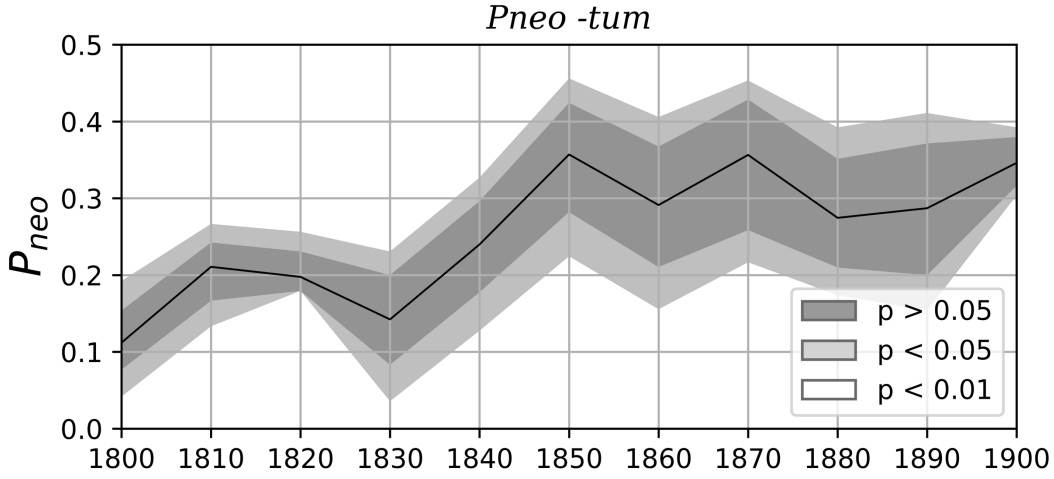

Figure 5: Mean P_{neo} values for -tum with with bounds ($p < 0.05$, $p < 0.01$) for 100.000 randomly resampled versions (with 4 million token per decade) of the corpus Deutsches Textarchiv (DTA).

For this pattern, there is a considerable gain in P_{neo} over the course of the 18th century. At the turn of the 19th century, roughly a third of all *-tum* types are new (cf. Berg submitted).

5. Conclusion

None of the above is intended to deny the merits of Säily's method. For measures of lexical richness in subcorpora determined by social rank, gender, profession and the like, it offers a robust and assumption-free tool, and the online publication of the data as SVGs is — in my opinion — the right way forward. But for measuring changes in productivity, new types are the better alternative. This method is more direct — in that it measures what it promised — and more directly interpretable.

Acknowledgments: I would like to thank Stefan Hartmann, Harald Baayen, and an anonymous reviewer for helpful comments on an earlier version of this paper.

References

Baayen, R. Harald. 1989. *A corpus-based approach to morphological productivity. Statistical Analysis and Psycholinguistic Interpretation*. PhD thesis, Free University, Amsterdam.

[8] Note that Säily's original method does not allow one to compute the significance of changes between different sub-periods, either: We can only check whether a subset of the data is significantly different from a larger set.

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In Geerd Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1991*, 109–149. Dordrecht: Kluwer Academic Publishers.

Baayen, R. Harald. 1993. On frequency, transparency, and productivity. In Geerd Booij and Jaap van Marle (eds.), *Yearbook of Morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. In Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics. An international handbook*, 900–919. Berlin: de Gruyter.

Berg, Kristian. submitted. Changes in the productivity of German word-formations patterns. Some methodological remarks.

Cowie, Claire and Christiane Dalton-Puffer. 2002. Diachronic Word-formation: theoretical and methodological considerations. In Javier E. Diaz Vera (ed.), *A Changing World of Words: Studies in English Historical Semantics and Lexis*, 410-436. The Netherlands: Rodopi.

Kempf, Luise. 2016. *Adjektivsuffixe in Konkurrenz. Wortbildungswandel vom Frühneuhochdeutschen zum Neuhochdeutschen*. Berlin: De Gruyter.

Klein, Wolfgang. 2013. Von Reichtum und Armut des deutschen Wortschatzes. In *Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache. Herausgegeben von der Deutschen Akademie für Sprache und Dichtung und der Union der deutschen Akademien der Wissenschaften*, 15–55. Berlin: de Gruyter.

Lenders, Winfried and Klaus-Peter Wegera. 1982. *Maschinelle Auswertung sprachhistorischer Quellen. Ein Bericht zur computerunterstützten Analyse der Flexionsmorphologie des Frühneuhochdeutschen*. Tübingen: Niemeyer.

Säily, Tanja. 2016. Sociolinguistic variation in morphological productivity in eighteenth-century English. *Corpus Linguistics and Linguistic Theory* 12(1). 129–151.

Säily, Tanja and Jukka Suomela. 2017. types2: exploring word-frequency differences in corpora. In Turo Hiltunen, Joe McVeigh and Tanja Säily (eds.), *Big and rich data in English corpus linguistics: methods and explorations*. (Studies in Variation, Contacts and Change in English 19). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/19/saily_suomela/