Changes in the productivity of word-formations patterns.
Some methodological remarks

Kristian Berg, Carl von Ossietzky
Institut für Germanistik
Universität Oldenburg
26111 Oldenburg, Germany
kristian.berg@uni-oldenburg.de

Accepted for publication in Linguistics

Acknowledgements: I would like to thank Frank Anshen, Mark Aronoff, Nanna Fuhrhop, Wolfgang Klein, Oliver Schallert, an anonymous reviewer, and particularly Harald Baayen and Stefan Hartmann for valuable comments on a first version of this paper.
Abstract

Changes in the productivity of word-formation patterns are often investigated using hapax legomena. In this paper, I argue that at least in diachronic investigations of productivity, a measure based on first attestations is a viable alternative to hapax-based measures. I show that such a measure is a more direct proxy to new words than hapax-based measures – it measures what we want to measure, which is not always true for the latter. I present a method that deals with the common problem of varying sub-corpus sizes (I suggest we randomly resample the sub-corpora up to a predefined size), and to the problem of old words appearing as new at the start of the corpus (I suggest we take an earlier corpus and determine a point in time when almost all old words have registered). Armed with these instruments, we can determine the ratio of new types to existing types for a time span, which can be regarded as the renewal rate of the respective category.

Keywords: Productivity, word-formation, diachronic morphology, new words

1. Introduction

What is morphological productivity and how can it be operationalized? This question has occupied morphological theory for decades. The common denominator is that productivity is based on new words. This holds for the locus classicus of productivity, Bolinger’s definition of “the statistically determinable readiness with which an element enters into new combinations” (Bolinger 1948:18, my emphasis), but it holds for many others as well, cf. e.g. Booij’s (2012:70) assertion that productive patterns “can be used to form new words”, or Bauer’s (2001: 98) definition of productivity as the potential of a morphological process “for repetitive non-creative morphological coining”. Baayen’s (1993:183) recapitulation still holds: “There is a broad consensus that productivity concerns the property of morphological processes to give rise to new words.”

From this conceptual foundation on new words arises the major dilemma for synchronic productivity: In a synchronic corpus, we cannot know which words are (or will be) new. This is where Baayen’s (1989, 1992, 1993) elegant idea comes into play: We may not know how many words are new, but we can use rare words as an indicator for those new words – a canary in the coal mine of sorts, as Mark Aronoff put it (p.c.). This works because productive patterns tend to have many low-frequent formations, whereas unproductive processes are characterized by more high-frequency types, cf. e.g. Aronoff (1983), Bauer (2001: 147).

The usefulness and the ease of application of Baayen’s measures, and the fact that they often correlate with linguists’ intuitions (cf. e.g. Baayen 1993), has made them the gold standard for synchronic productivity. They are also attractive for diachronic investigations of productivity, cf. (among others) Scherer (2005), Štichauer (2009), and Trips (2009). It is crucial, however, to remind ourselves that “P is an indirect measure of productivity” (Bauer 2001: 153). There is a much more direct measure when dealing with diachronic data: New words (cf. e.g. Cowie 1999, Anderson 2000). It is easy to determine for each time span in a diachronic corpus which words are new. There are some methodological challenges; addressing them is one of the purposes of this paper. The quintessential point is this: Productivity is conceptually based on new words, and we can determine new words in a diachronic corpus. A measure based on new words has a number of other advantages, as I will show below.
Like most research on morphological productivity, this investigation will be limited to
derivational word formation. I use five German derivational suffixes to demonstrate the
usefulness of the proposed measures.

Of course, there are other aspects of productivity that a focus on new words does not
capture. For example, both the number of types and the number of tokens in use at any
given time highlight important aspects of a pattern’s morphological productivity (see e.g.
Baayen 2009 for an intuitive commercial analogy). Those aspects are largely neglected in the
present paper as measures in their own right. This is simply due to space limitations, not
because I believe them to be unimportant.

The structure of this paper is as follows: In the next section (2.), I will discuss related work on
diachronic productivity, before I give an overview of the corpus I have used for the
investigation and the derivational suffixes I have looked at (3.). After that, I will give a brief
overview of approaches to diachronic word formation, and building on this, I will develop a
measure of diachronic productivity based on new words (4.). I will then use this measure in
an investigation of five German suffixes (5.). The measure is compared to the most widely
used measure in diachronic productivity, Baayen’s $\mathcal{P}$, in the subsequent section (6.). The
main findings are summarized and discussed in the last section (7.).

2. Related work: Diachronic productivity

The field of research on productivity is vast. In the following, I will limit the discussion to
diachronic productivity, and more precisely, to quantitative approaches. These approaches
basically come in two different flavors: They are either based on new formations, or they use
Baayen’s hapax-centered measures. I will comment on both in turn.

Approaches that measure productivity using new words rely mostly on dictionaries. One of
the first works of this kind is Neuhaus (1971, 1973), who measures the number of first
attestations of -ish words in the OED per decade. In a similar manner, Anshen & Aronoff
(1989) use a reverse dictionary (Walker’s rhyming dictionary) to collect words with a given
suffix (in this case, e.g. -ness and -ity). These words are then checked for the date of their
first attestation in the Oxford English Dictionary (OED). This way, they can determine which
pattern produces more new words. Arguably, the greatest weakness of this approach is that
only those words with an entry in the reverse dictionary can be checked for their first
attestation. Ten years later, Anshen & Aronoff (1999) manage to overcome this limitation by
using a more sophisticated method on an electronic version of the OED. They semi-
automatically extract every word in the OED with -ment and -ity to determine the factors for
the decline in productivity of the former and the stability of the latter. Anshen & Aronoff
essentially use the raw numbers of new words per time span (in their case, half-century) and
compare the numbers for the two suffixes: Whichever has more new words per time span is

\[\text{\textsuperscript{1}}\text{ For a comprehensive general overview of productivity, see e.g. Bauer (2001) or Baayen (2009); for a}
\text{comprehensive overview of productivity measures see Hilpert (2013: 127ff.); for a general overview of}
diachronic productivity see Scherer (2015).}
\text{\textsuperscript{2}}\text{ There are also approaches like Plag (1999) and Bolozky (1999) that use measures based on new words as well}
as those based on hapaxes.
more productive. Note that this method actually only allows Anshen & Aronoff to compare
the productivity of different patterns in the same time span. Statements like “the number of
borrowed -ity forms slowly increases and then, starting at the second half of the 17th
century, decreases” (Anshen & Aronoff 1999: 21) are not wrong, but they cannot be
generalized to mean that the productivity declined from 1750 onwards: It may simply be
that the OED contains less data in that period. To avoid this problem, the data need to be
normalized. This is a notorious issue in quantitative diachronic linguistics, and I will return to
it several times in this paper. The question of what to normalize for is far from trivial. Lindsay
& Aronoff (2013), for example, building on the method used in Anshen & Aronoff (1999),
decide to use the number of types in the OED for each time span.

According to Baayen & Renouf (1996: 69f.), there is a fundamental problem with the use of
dictionaries. Truly productively formed words, which are semantically transparent and very
rare, will often not be included in the lexicon, primarily for economic reasons. As a
consequence, dictionary-based measures underestimate the actual productivity, at least of
productive processes (but cf. Plag 1999: 97-99 who argues that arguments like this against
the use of dictionaries do not apply to the OED).

Dalton-Puffer (1996), Cowie (1999) and Cowie & Dalton-Puffer (2002) avoid the pitfalls of
dictionaries by counting new words in diachronic corpora. The problem of normalization
remains, though: If the corpus is not evenly sized, that is, if certain time-spans consist of
more running words than others, what should the count data be normalized for? Cowie &
Dalton-Puffer (2002: 427) argue against the use of tokens – as in e.g., ‘In 1800, there are 31
new -isch types per 1,000,000 tokens’ – because that would amount to “not counting like
out of like”. Yet then the authors go on to do just that, and normalize new words for 100,000
tokens (p. 430f.). I will return to this problem below (section 4).

Hapax-based measures, on the other hand, have been developed by Baayen (1989, 1992,
1993) for synchronic investigations. As briefly noted above, they provide an elegant way to
gauge the synchronic productivity of a word formation pattern without access to
information about which words are new. The central assumption is that productive patterns
produce many low-frequency words. The reasoning for this is as follows: Low frequency
words, especially those with just one attestation, so-called hapax legomena, are probably
semantically transparent; if they were not, their meaning could hardly be inferred.

Productive patterns, on the other hand, produce semantically transparent words; it thus
seems probable for productive processes to produce many rare formations. The most widely
used measure $\mathcal{P}$ is called ‘productivity in the strict sense’ (Baayen & Lieber 1991: 817),
‘potential productivity’ (Baayen 2009), or ‘category conditioned degree of productivity’
(Baayen 2001:157). It relates the number of hapaxes of a category to the total token count
of the category ($\mathcal{P} = \frac{n_1}{N}$ with $n_1=$hapaxes and $N=$sum of all tokens). As such, $\mathcal{P}$ is the
probability of encountering a hapax after N tokens.

The ease of computation and the fact that the measure yields intuitive results has led to the
acceptance of $\mathcal{P}$ as the de-facto standard for measuring productivity. It is not surprising that
it is regularly used as the central measure in diachronic investigations as well, cf. e.g. Scherer
(2005), Schneider-Wiejowski (2011), Trips (2009), or Schröder (2011). The way the measure
is used is that the corpus is divided into consecutive sub-corpora, with $\mathcal{P}$ values computed
for each sub-corpus. There are two main problems with this. First, the corpus has to have a
certain minimal size: In smaller corpora, there will be more hapaxes. To put it differently, many words that are not really rare will nevertheless appear only once in a small corpus simply because it is not large enough to include the respective second attestations (cf. Dal 2003: 17). Second, the pitfalls discussed above apply: To be able to compare two \( P \) values, the sub-corpora should be of equal size, or they should be normalized. Otherwise, differences between the values are much harder to interpret; they may simply be due to one time span accidentally containing more tokens than the other. However, as the relation between tokens and hapaxes is not linear (cf. Dal 2003: 17), any linear normalization is necessarily too simplistic. A solution to this problem is to extrapolate the number of hapaxes for sample sizes larger than the actual size (cf. e.g. the model of Baroni & Evert 2005). And as Gaeta & Ricca (2006) have argued, it is not the sample size of the (sub-)corpus that needs to be standardized, but the token count for individual patterns. The larger the number of tokens of a pattern, the lower is \( P \). If a pattern is instantiated by many tokens, its \( P \) value is automatically lower than that of an infrequent pattern. Comparing the unadjusted \( P \) values will thus “always imply an overestimation of the values of \( P \) for the less frequent suffixes”, as Gaeta & Ricca (2006: 63) state. Accordingly, contemporary investigations of diachronic productivity (e.g. Schneider-Wiejowski 2011, Hartmann 2018) still center around Baayen’s \( P \), but they modify their corpora with elaborate mathematical extrapolation procedures like finite Zipf-Mandelbrot models.

3. Data bases and suffixes

Before we turn to the methods I would like to put forward, let us take a brief look at the data that will be used to test these methods. I will use the large diachronic corpus Deutsches Textarchiv (DTA) as the main data base in this paper (cf. www.deutschestextarchiv.de). As of December, 2017, it contains 2,419 printed texts from the late 15th century to the early 20th century, totaling 138 million tokens. It is not balanced for genre; it includes newspaper texts, scientific publications, novels, pamphlets, etc. The main advantage of the corpus, however, and the primary reason for its particular useability for a diachronic investigation like this one, is that it is graphemically normalized and lemmatized. The original spelling of every word is accessible, but it is enriched with information about graphemic and lexical invariants: Each running word is mapped onto a graphemically normalized token form (e.g., the spelling variants Kleid, Kleidt, Kleyd, and Kleydt ‘dress’ are all mapped onto the same invariant form Kleid). Each graphemically normalized token is then mapped onto one lemma (e.g. the normalized tokens Kleid ‘dress’ and Kleider ‘dresses’ are mapped onto the lemma Kleid). With this lemmatization and graphemic normalization, it is possible to easily extract all lemmas that end with a certain suffix.

Reasons of validation (see next section) make it necessary to also use corpora that cover the time before and after the main corpus. To cover the time before the DTA corpus, I chose the Early High German (EHG) corpus Bonner Frühneuhochdeutschkorpus (http://www.korpora.org/fnhd/, Lenders & Wegera 1982). This corpus contains 40 Early New High German texts from the 14th to the 17th century with a total of around 600,000 tokens. Like the DTA corpus, it is lemmatized, which facilitates the extraction of types. The time following the main corpus is covered by the 20th century core corpus of the Digitales
Wörterbuch der Deutschen Sprache (www.dwds.de). It contains between 10 and 13 million running words per decade and is balanced for genre (cf. Geyken 2007). Even though the corpus is not fully accessible due to copyright laws (it cannot be downloaded, for example, and only a limited number of results can be displayed), it has an application programming interface (API) which is used to automatically collect frequency data from the web server of the corpus.

The suffixes I investigate are all derivational suffixes from different domains, adjective suffixes (-isch, -bar), nominal suffixes (-nis, -tum), and a verbal suffix (-ieren). The suffixes exhibit different degrees of productivity: -isch and -bar are supposedly highly productive in modern-day German (see below for references); -nis and -tum, on the other hand, are regarded as hardly productive, and the status of -ieren is unclear. For one suffix, -isch, there is a recent investigation concerning its diachronic productivity (Kempf 2016); this is not the case for the other suffixes. Accordingly, the following short summary will mostly refer to the kind of base the suffix operates on, the semantic relation between base and derivative, and the synchronic productivity.

-isch: This pattern yields adjectives from nouns, which are mostly native person nouns (e.g. mörderisch ‘murderous’). It also serves to integrate bound non-native bases as adjectives (e.g. elektrisch ‘electric’). In her diachronic investigation, Kempf (2016: 248ff.) observes a strong increase in productivity between 1450 and 1700 (using a variety of hapax- and neologism-based measures). Kempf hints at a slight decline in the 19th century. Today, the pattern is regarded as highly productive (cf. e.g. Lohde 2006: 184).

-bar: This is an adjective suffix that operates mostly on verbal bases. These verbal bases are in turn mostly transitive verbs that form a regular passive (cf. Fleischer & Barz 2012: 332ff.). The resulting adjectives have a passive-like quality and denote the possibility of the action the verb describes (e.g. essbar ‘able to be eaten’). Today, the pattern is the most productive pattern for de-verbal adjectives (Fleischer & Barz 2012: 332). Historically, the suffix also operated on nominal bases, e.g. dankbar ‘thankful’, cf. Flury (1964).

-nis: This suffix forms abstract nouns from verbs (e.g. Besäufnis ‘the act of getting drunk’). Nominal and adjectival bases are very rare. The pattern is unproductive today (cf. Motsch 2004: 359).

-tum: This noun suffix operates mostly on nominal bases, more precisely, on person nouns. It yields abstract nouns that may denote both the key qualities of the person (or group of persons) that the base refers to (as in e.g. Abenteuerertum ‘the quality of being an adventurer’) or the collectivity of those persons (as in e.g. Bürgertum ‘the collection of all citizens [i.e., the bourgeoisie]’). This pattern is only marginally productive today (cf. Lohde 2006: 107).

-ieren: This verb suffix operates mostly on foreign bases; many of them are bound bases (e.g. informieren ‘to inform’), but there are a few native free bases as well (e.g. amtieren ‘to hold office’). If the bases are free, they are mostly nouns. Semantically, the formations are very heterogenous (cf. Fleischer & Barz 2012: 432).
4. Methodology

The method proposed here is reasonably straightforward: Divide a diachronic corpus into time-spans and count the new words with a given word formation pattern. As with any apparently straightforward measure, the devil is in the detail: What is an appropriate time span, and which words belong to the word formation pattern in question? The size of the time span hinges on the corpus size, but is apart from that largely a matter of convenience.

From 1700 on, the DTA contains for the most part at least 4 million tokens per decade (data before 1700 will be discarded for reasons independent of periodization, see below). A minimum of 4 million words per decade sub-corpus seems to be a reasonable size compared to much smaller synchronic corpora like the 600,000 token Eindhoven corpus, for which Baayen (1993:187) claims that “the major patterns of productivity already emerge very clearly”.

Which words belong to the word formation pattern in question? Is e.g. Nutzbarkeit ‘useability’ an instance of the word formation pattern -bar (nuttbar ‘useable’), which in turn derived with the nominal suffix -keit? I argue with (among others) Baayen (2009: 903) that only those words should be included where “a given rule was the last to apply” (in a synchronic-structural sense, not historically). This has the advantage that the extraction of the pattern from the corpus is greatly facilitated: We simply search for lemmas that end with the suffix in question and meet the part of speech criteria\(^3\). To avoid results like frisch ‘fresh for -isch, the search pattern demands that the suffix be preceded by at least three letters; there are only a few German stems with less than three letters (cf. Berg 2017: 115ff.). The results are then corrected and cleaned manually. In some instances, a faulty lemmatization leads to two distinct entries for one lemma (e.g. Kaisertum, Keysertum ‘empire’); this is adjusted. In other cases, the pattern in question is not the last step in the derivation: for example, undruckbar ‘unprintable’ has to be analyzed as a derivation of the adjective druckbar ‘printable’, and the entry thus has to be excluded. The same holds for the nominal suffixes, where compounds with -tum and -nis as the second element has to be filtered (e.g. Familienheiligtum ‘family sanctuary’). And finally, many Latin words (or words of Latin origin) ending in the letter sequence nis or tum are excluded (e.g. tyrannis, Quantum ‘quantum’). Entries without free bases, however (like e.g. elektrisch ‘electric’), and semantically opaque entries are not excluded (contra Gaeta & Ricca 2006:74f.). These entries still conform to the output constraints of the word formation pattern (cf. e.g. Aronoff 1976).

Apart from these questions, there is a general problem that is an almost chronic issue for diachronic investigations: Many diachronic corpora vary in size for their time spans. As a case in point, take the corpus used in this paper. Figure 1 shows that the decades from 1470 to 1900 vary greatly with respect to the tokens they contain. This holds both for the sum of tokens in each decade as well as for the tokens in each genre.

\(^3\) While this catches most of the word forms we are looking for, some still escape because of faulty lemmatization. An additional token-based search for inflected forms and spelling variants is necessary (e.g. for -tum I searched for tokens that end with either t(h)um, t(h)um(e)s, t(h)ümer, t(h)ume or t(h)ümern (and that have not already been collected).
Not surprisingly, more recent decades contain more data (although there are exceptions). For a productivity measure based on new words, this is problematic because more tokens mean more types – but, as mentioned above, the relation is not linear. Thus we cannot simply list the set of new words for each decade if the decades differ in size as considerably as they do in the DTA, and we can also not normalize for tokens and state the number of new words relative to some chosen number of tokens (e.g., ‘in 1710, there are 14 new words with -isch per 1,000,000 tokens’).

The easiest way to deal with this issue is to look at the number of new types as a function of the number of tokens in that particular time slice. Figure 2 plots this relationship for the five suffixes. The y-axis could equally well be labelled ‘vocabulary’; the graph shows a diachronic vocabulary growth curve in the sense of Baayen (1992: 113). In some sense, this is the ‘purest’ visualization of the data; it does not exclude any data and rests on a minimum of presuppositions. Each dot represents the number of types of a word formation pattern in a given decade.

*Figure 1: Barplot of the number of tokens per decade in the corpus Deutsches Textarchiv.*
Figure 2 reveals several facets of the data. First, the size of the decades and centuries (in tokens) is visible as the horizontal distance between the dots. It varies with respect to the tokens they contain: the 19th century has more tokens than the 18th, which in turn has more than the 17th.

Second, the size of the combined types at the end of the corpus varies: -isch has the largest vocabulary over time (4,971 types), and -nis the smallest (183 types). Take a moment to note, though, that the number of types at the end of the corpus is not the number of types in the last decade; it is the accumulated set of types that was gathered over time, and it contains words that have fallen out of use, and many words that were used just once a long time ago. The last substantial decade in the corpus, 1890–1899, contains 2,088 different types with -isch (as compared to the 4,971 types in the whole corpus), and 65 different types with -nis (as compared to the 183 types in the whole corpus). Baayen (2009: 901f.) is thus right in his assessment that the size of the vocabulary of a word formation pattern (what he calls “realized productivity”) is – at least partly – a “past achievement”; it contains the successful word formations from the past, as well as new ad-hoc formation; but it does not contain past ad-hoc formations. Even though realized productivity is part of a synchronic investigation of productivity, it is a rather coarse measure for a diachronic investigation. We can disentangle past and present achievements for each time span by using first attestations, and thus arrive at a more accurate measure, as I will show below.

Third and most importantly, figure 2 shows the rates with which new types appear, and how this rate changes over time. For example, new -isch words are produced at a higher rate than new -ieren words from around 1750; the -isch curve rises more steeply than the curve for -ieren. The slope of the line thus indicates changes in productivity. This slope of the line is the difference in types divided by the difference in tokens, $\Delta$new types/$\Delta$tokens (if we used approximation functions to the lines in figure 2, we would look at their first derivative, cf. e.g. Baayen 2001: 49). Note that the difference in types in each time span equals the newly attested words. The result for the five suffixes is plotted in figure 3:
Figure 3: Change in the number of new words with a given pattern divided by the change in tokens with that pattern for the five patterns (-bar, -isch, -nis, -ieren, and -tum). Data base: corpus Deutsches Textarchiv (DTA).

In this representation, we can clearly see the level of productivity and its changes over time. The suffix -isch has the highest rate with around 0.00003 new types per new token, or 30 new types per 1,000,000 tokens. The level is rising slightly over time. The level for -ieren is stable (at least from around 1750 onwards) at a considerably lower rate of below 0.00001 new types per new token. The level for -bar slowly rises to 0.00001 new types per new token. For -tum, the rise starts later (around 1800), and it does not reach the level of -bar. Finally, there do not seem to be any changes in productivity for -nis – the word formation pattern is unproductive.

There are three major problems with this kind of methodology. One is that the curves in figure 3 all show a bias towards higher productivity at earlier stages. A second one is that we are de facto normalizing over tokens, even though, as was stated above, the relation between types and tokens is not linear. And finally, if we are calculating new types as a function of tokens, we are “not counting like out of like” (Cowie & Dalton-Puffer 2002:427). I will address these issues in turn.

The bias towards higher productivity levels at the beginning of the corpus has often been observed (cf. e.g. Cowie & Dalton-Puffer 2002: 429; Kempf 2016: 116). It follows from the fact that words which predate the corpus “must initially register” (Cowie & 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. There are several ways to deal with this issue. Cowie (1999) suggests
using a different, earlier corpus as a ‘starting lexicon’: If any given word already appears in
the older corpus, it does not count as a new word. While this is generally a good idea, a
prerequisite is that the older corpus is large enough to actually contain all (or most) words
that predate the later corpus.

I would like to present a different way to handle this problem, one that also works if the
older corpus is considerably smaller than the actual corpus (for a similar but more qualitative
approach, see Kempf’s (2016) and Kempf & Hartmann’s (2018) "comparative dating
method"). The idea is simple: We determine the set of types in the earlier corpus and track
their first attestation in the actual corpus. This should lead to a curve which is high at initial
stages of the newer corpus (older words are registering), but which should flatten at some
point – and it is this point (or period of time) that we are interested in. From here on, the
majority of older words has been encountered.4

For the present investigation, I used the Early New High German corpus introduced above
(section 3). This corpus overlaps with the DTA corpus, which starts in 1473; as a
consequence, I only use texts up until 1500 from the Early New High German corpus. From
these 20 texts, a set of 4,919 distinct types is extracted, and the first attestation of each of
these types in the DTA is determined.

Of these 4,919 types in the Early New High German corpus, 3,789 also appear in the DTA
corpus (77%). As expected, most of them are first attested in the early stages of the DTA.
Figure 4 shows a barplot of the number of types according to the decade of their first
attestation in the DTA:

---

4 Another method that could even do without an older corpus is to use morphologically simple words as a
baseline and track their vocabulary growth curve. Morphologically simple words predate cum grano salis
today’s corpora. Once most of these morphologically simple words have appeared (i.e., once the growth curve
flattens), we can infer that most older words in general have registered. If the corpus is not annotated
morphologically, we could use short words as an approximation to simple words, e.g. the set of four letter
words (I would like to thank Mark Lindsay for this idea).
Words from the earlier corpus are first attested even in the latest decades of the DTA. But this is mostly due to the lemmatization practice of the Early New High German corpus: For example, the token *paraliticus*, which appears twice in the corpus, is clearly a Latin loan word (evidenced by the Latin suffix -us); yet it is lemmatized as *Paralytiker*, which first appears in the DTA corpus around 1900. Thus the actual distribution is probably even clearer, with fewer types at later stages. In any case, 1700 is a sensible cut-off point: As can be seen in figure 4, the number of newly attested (old) types after this decade is marginal. In total, only 140 of the 3026 types from the Early High German corpus are first attested 1700 or later; that is a ratio of 4.9%. That means that the DTA data before 1700 is potentially (though not necessarily) skewed when it comes to first attestations, and we should focus on the decades from 1700 on.

The second problem with figure 3 mentioned above is that we are normalizing the token count – we arrive at statements like ‘for *-isch*, there are about 30 new types for every 1,000,000 tokens’. But again, the relation between types and tokens is not linear, and we have every reason to believe that this also holds for the relation between new types and tokens. The remedy I suggest is twofold: Make the time spans similar with regard to their token count, and normalize for types instead of tokens. Both steps lead to a measure that is at the same time more stable and a more direct proxy to new words, as I will show below. This measure also takes care of the third issue raised above, the fact that we were counting types out of tokens.

For corpora that are not balanced over time (as the DTA, see figure 1 above), adjusting the size of the time spans is the easiest way to ensure comparability of different time spans. The first step is to determine the target size of the time span. As mentioned above, the minimum size almost all decades after 1700 have is 4,000,000 tokens (remember that we are excluding data before 1700), so we aim for this size in each decade. Instead of carefully curating texts for the decades (which is of course possible), I suggest we use a Monte Carlo simulation: We randomly re-sample each decade’s texts out of all texts for this decade until we reach the target size, then we move on to the next decade until all decades have been re-sampled. We do this 100,000 times, and we thus end up with 100,000 alternatively re-sampled versions of the main corpus. We note for each permutation of the corpus the number of new types and the number of all types with the word formation pattern in question. After all permutations have been computed, we calculate the mean and the range into which 95% and 99% of the data fall.

If the time spans are of equal size, we can simply count and compare the numbers of new words in each decade without normalizing. I will call this measure $V_{\text{neo}}$. Suppose pattern A has 200 new words in a given decade, and pattern B has 100 new words in the same decade, then pattern A has a higher $V_{\text{neo}}$ value; it is more productive according to this measure.

However, comparing the changes in new words for two different word formation patterns is potentially difficult if their respective vocabulary sizes differ. Using the example above, suppose the 200 new words for pattern A occur alongside 200 old words, while the 100 new

---

5 More precisely, we add the decade’s texts in random order until we exceed the target size; for the last text, we cut the excess tokens so that the total number of tokens matches the target size. By keeping the texts intact, we accommodate for the fact that words are not distributed randomly throughout texts and collections of texts (cf. Säily 2016: 132).
pattern B words occur with just 50 old words. This changes the picture: Pattern B has a higher renewal rate, and we should take that into account. This renewal rate is the same as Bolozky’s (1999: 5) R1, the “pattern’s potential to ‘regenerate itself’”, and it is similar in spirit to Baayen’s (1993) category conditioned degree of productivity\(^6\). The reasonably simple formula for the neologism-based productivity \(P_{\text{neo}}\) for a time span \(t\), a corpus size \(N\), and a word formation pattern \(wfp\) is given in (1). We count all types that first occur in \(t\) and divide them by the count of all types attested in \(t\):

\[
(1) \quad P_{\text{neo}} (wfp, t, N) = \frac{\text{new types (wfp} t, N\text{)}}{\text{all types (wfp} t, N\text{)}}
\]

\(P_{\text{neo}}\) can be interpreted as a probability, the probability of encountering new types with this pattern; the \(V_{\text{neo}}\) measure introduced above is simply a count measure. The main advantage of \(P_{\text{neo}}\) is that it is directly interpretable as the fraction of all types in a time span that are new. To give an example from the actual investigation (see below), in 1840–1849, there are 1,372 types with the pattern -isch in the first of 100,000 4-million token versions of the DTA corpus, and 250 of them had never occurred before (i.e., in the previous 4-million word samples leading up to 1840). That means 18.2\% of the types in that decade were new.

Compare that to a pattern which is usually believed to be unproductive, -nis: In the same decade, only three of the 68 total types were new, which is a ratio of 4.4\%.

Baayen’s \(P\), on the other hand, denotes the fraction of all tokens with a word formation pattern that occur only once. It rests on the assumption that hapaxes are indicative of new words. For diachronic investigations, the productivity measure introduced in (1) is potentially more useful: First, it is more direct in that it rests on first attestations, which are a close approximation of ‘actual’ new words in the language community; we do not need to take the (synchronously necessary) detour via hapax legomena. Second, the new words can in turn be investigated qualitatively: How did the new formations change, and what does that say about the respective constraints of the word formation process? And third, if we tie productivity to the ‘birth’ of words, we can also take the ‘death’ of words into account and look at the life-span of word formations. Are they nonce formations, or do they become institutionalized or lexicalized? And the measure in (1) is not only more useful in diachronic investigations, it is also more stable than hapax-based measures. I will demonstrate this in the following sections.\(^7\)

A few methodological issues need to be addressed. When is a word a neologism? This seems to be a trivial question, but there are several pitfalls. First of all, we can only ascribe this status to words relative to the corpus we are using. In a larger corpus, we will probably find

---

\(^6\) Cowie (1999: 133f.) uses the same measure, although without keeping the time spans at a fixed size (and without giving it a name). I would like to thank Stefan Hartmann for pointing this out.

\(^7\) There is one caveat: We are only making statements for the productivity of the patterns in a corpus of size \(X\), and a time span of size \(Y\). Yet with this in mind, we can very well compare patterns that differ in their total token count. Gaeta & Ricca (2006) argue that Baayen’s \(P\) measure is systematically too high for low-frequency patterns, and they suggest to calculate the expected number of hapaxes for larger token counts. This is generally not necessary if we remind ourselves that the productivity values we obtain are always only for certain corpus sizes. After all, that is psycholinguistically plausible: Language users gauge the productivity of a word formation pattern on the basis of the input they get – and that input (in the form of running words of text) is the same for all word-formation patterns, which in turn means that some (like compounding in German) are much more frequent than others (for example derivation with -nis).
earlier attestations for many words. If we accept that, then a word in the corpus is new if it was never attested before in the corpus. As a corollary of this definition, words that appear again after a longer period of absence are not new. Take the adjective *katalanisch* ‘Catalan’, for example. It is first attested in the DTA corpus 1653, where it appears once; the next attestation is in 1821, after 168 years of radio silence. In a certain way, the 1821 attestation is also new: The writer had probably never seen the word *katalanisch* in print before, and one can argue that the word was thus new in the 1821 document. But this distribution may as well be an effect of the corpus size and the low frequency of the word itself; after all, we have to remind ourselves: We are only monitoring a small sample of language use over time. It may well be that the large gap we witness is accidental. For this reason, we adhere to the rather strict definition given above: A word is new when it is first attested; from then on, it is old, no matter for how long it is not recorded after that first attestation.\(^8\)

Second, and more general, the approach to newness is very naïve. All words that occur for the first time at some point in the corpus are new, but of course, at least some of them were coined deliberately and intentionally, for example by writers (I would like to thank Harald Baayen for pointing this out).

Third, quantitative approaches cannot readily capture changes in the function of the suffix. Take the German suffix *-mäsig* as in *wettermäsig* ‘weather-wise’. The Old High German version of this suffix, *-māzi*-māzig referred more explicitly to the base *māza* ‘measure’: *Xmāzi* meant ‘having the dimension of X’ (Fleischer & Barz 2012: 346).

Fourth, it is well documented that productivity in word formation also varies with register or genre (cf. e.g. Plag, Dalton-Puffer & Baayen 1999). Even though the DTA corpus is annotated for genre, it is not balanced for genre, and the contributions of the genres to the overall token count vary considerably, making it difficult to use the sampling method introduced above. Accordingly, in what follows I will ignore the influence of genre – knowing full well that I am probably discarding information that could account for some variation in the data.

And finally, using a fixed target size of tokens for all decades, we run into the conceptual problem that the historic population of the German speaking countries grew over time (which in turn lead to a steadily growing output of texts over time), although not monotonously (cf. the Thirty Years’ War). Cutting each decade at a fixed target size ignores the link to the development in population (and text production), as Harald Baayen observes. However, this link is only implicit in the DTA corpus anyway: it strives neither for complete coverage of available texts, nor is the number of texts in the corpus proportional to population growth or to the number of published texts. For this reason, I will use the methodology sketched above and limit each decade to a fixed target size. The conceptual problem remains, of course – I simply cannot address it with the data at hand.

---

\(^8\) This is, as Harald Baayen rightly observes, a cumulative societal perspective on the lexicon. For other perspectives, other measures might be more appropriate. For example, to describe the knowledge that individual speakers have of a word formation pattern, the type count may be more suitable: neologisms tend to be rare and – depending on the amount a person reads – may not ‘reach’ that person at all.
5. Case study: German derivational suffixes

Let us apply the measures developed in the last section to the five word formation patterns in the 4 million token versions of the DTA corpus. As mentioned above, we are not dealing with one version of the corpus, but with 100,000 randomly sampled corpora. As a consequence, we can compute a mean value of new words, but we can also compute the probability distribution of new words: We can state the probability with which any new resampling of the corpus will fall above or below certain limits. These limits are given for the p < 0.01 level (light grey area) and the p < 0.05 level (dark grey area) in the following plots.

Figure 5 shows the $V_{\text{neo}}$ values for the five suffixes:

![Graphs showing $V_{\text{neo}}$ values for five German word formation patterns (-tum, -bar, -ieren, -isch, and -nis).](image)

The pattern -isch produces the highest number of new words per decade, and the amount is slightly rising. The patterns -ieren and -bar yield less new words, but they are in the same range, with -bar producing slightly more words, and the pattern also rising. The number of new -tum words increases as well, but the numbers are well below those for the other suffixes, except for -nis: This seems to be a truly unproductive pattern in the period investigated. It would be premature, though, to reduce productivity to the $V_{\text{neo}}$ values: After all, there are many more -isch types (not just the new ones) than e.g. -tum types. We should thus take the size of the morphological category into account, as explained above. Doing this yields the following $P_{\text{neo}}$ values:

---

9 The dataset and the Python scripts are available at https://github.com/kristian-berg/Linguistics.
Figure 6: $P_{\text{neo}}$ values for five German word formation patterns (-tum, -bar, -ieren, -isch, and -nis); light grey area: 99% confidence intervals, dark grey area: 95% confidence intervals. Data base: 100,000 times randomly resampled 4 million token version of the corpus Deutsches Textarchiv (DTA).

One of the most visible differences between the plots is in the confidence intervals: For -isch and -ieren, they are comparatively small, while they are relatively large for e.g. -tum. This mirrors the difference in the number of types (and new types) between the patterns in the raw (i.e., unsampled) data: While there are, for example, 947 -isch types in 1800 (153 of which are new), there are only 28 -tum types (two of which are new). The probability for variation due to different samples is thus much higher for patterns with a lower type count. These results have to be handled with care; however, we know in which area the next sample will fall with a 95% or a 99% chance.

Turning back to the plots, we can observe the following: The adjective suffix -isch has the steadiest distribution. The $P_{\text{neo}}$ value is constant around 0.1 for the 19th century, with slightly higher values in the 18th century (this corroborates Kempf’s findings about the productivity of -isch on the basis of type counts, token counts, hapaxes and neologisms, cf. Kempf 2016:248ff.). That means this word formation pattern had a constant renewal rate of around 10% for a long time: In each decade, every tenth word with this suffix was new.

As an effect of this constant influx of new words, both the type and token counts of this pattern rise over time; the pattern is expanding because the productivity is stable at a certain non-negligible level. Figure 7 shows the number of -isch types and tokens over time:
As is visible, an apparently moderate renewal rate of 10% is mirrored as a considerable and almost linear increase in both type and token frequency for -isch over time. A constant $P_{\text{neo}}$ value leads to a constant rise in types and tokens. It would be interesting to see whether this generalization holds beyond the data presented here.

The second pattern in figure 6 above, -ieren, is also constant, but at a lower level of around 0.05–0.1. In contrast to that, the patterns -bar and -tum both show rising $P_{\text{neo}}$ values over the course of the 19th century. $P_{\text{neo}}$ for -bar even reaches the 0.3–0.4 mark, which means that roughly a third of all types with this suffix in the respective decades were new. Of the patterns investigated here, -tum and -bar are the most productive ones in the 19th century.

For -tum, the rise of productivity in the 19th century is mirrored in a rapid increase in -tum types in the 19th century: The lexical richness triples between 1800 and 1890 (figure 8, left panel).
This effect is not mirrored in the token count, however (figure 8, right panel), which –
despite occasional peaks – rises only slightly. One explanation for this distribution is that a
lot of the new words are rare and carry no weight compared to the established, high-
frequency types. For example, in 1800–1809, the three most frequent types Reichtum
‘riches’, Eigentum ‘property’, and Irrtum ‘error’ account for more than half of the total -tum
tokens of that decade (860 of 1,372 tokens, 63%). The situation has changed only slightly by
1890, when the three most frequent types, Eigentum, Christentum ‘christendom’, and
Reichtum account for less than half the tokens (2,094 of 4,935, 42%). As a contrast, the three
most frequent -isch types in 1800 (politisch ‘political’, französisch ‘French’, and römisch
‘Roman’) account for only around 9% of all tokens in that decade (1,411 of 16,589). The -tum
pattern is thus much more dominated by high frequency types than the -isch pattern.

The last word formation pattern in figure 6, -nis, is only marginally productive. In five
decades between 1700 and 1890, the productivity rises to around 0.1, but in the remaining
decades, the productivity ranges between 0 and 0.05, which means that at least 95% of the
types in the corresponding decades are old. What is more, there are only comparatively few
types overall: In any given decade, there are hardly more than 60 types (figure 9, left plot):
However, as witnessed in the \( P_{\text{neo}} \) plot above, the pattern is not totally unproductive; the vocabulary increases, albeit slowly compared to the other patterns. The token count increases even more slowly (figure 9, right plot). It is rather high considering the limited number of types: This is due to a number of high frequency types; for example, the three most frequent \(-\text{nis}\) words in 1800, \textit{Verhältnis} ‘relationship’, \textit{Kenntnis} ‘knowledge’, and \textit{Bedürfnis}, ‘need’ account for 59% of the total token count in that decade (3,106 out of 5,287 tokens). The fact that there are new words with \(-\text{nis}\) at all is surprising given both measures, the small vocabulary size and the relatively high token count. In cognitive models of word formation, a large vocabulary size is often seen as a facilitating factor for productivity, and the same holds for a large number of low-frequency words (cf. e.g. Taylor 2002); both are not the case for \(-\text{nis}\). The growth of the pattern is even more surprising when we look at the frequency of the new formations: Many of them become lexicalized or institutionalized, i.e., they are in a sense very successful word formations. This may be the key to understanding the marginal productivity of this pattern: There was an onomastic demand. I will come back to this point shortly.

One of the advantages of the measure used here is that we can take the set of neologisms in each time period as a starting point for a qualitative analysis, much like e.g. Kempf (2016) or Hartmann (2016). Changes in productivity can be caused by changes in onomastic demand (cf. e.g. Scherer 2007: 268 for German \(-\text{er}\) ); sometimes, the productivity of a pattern changes with the productivity of a competing pattern (cf. e.g. Lindsay & Aronoff 2013 for English \(-\text{ity} \) vs. \(-\text{ment}\) ); sometimes a pattern seems to fall out of fashion (cf. e.g. Berg submitted for German \(-\text{n}i\)-). In all of these cases, an analysis based on the set of neologisms per time period can shed more light on the language-internal (competition) or external (onomastic demand, changing fashions) factors. Changes in productivity also sometimes go hand in hand with changes in the input restrictions of a word formation pattern. As a pattern becomes more productive, the set of possible bases widens. Turned into a deterministic rule (productivity as inversely proportional to the amount of input restrictions, e.g. Booij 1977: 5), this is clearly wrong: There is a variety of reasons why patterns become productive or

Figure 9: Number of types (left panel) and tokens (right panel) with \(-\text{nis}\) over time; light grey area: 99% confidence intervals, dark grey area: 95% confidence intervals.
stop being productive. Still, both may be related, and a qualitative analysis can uncover this relationship.

In the following, I will sketch such a qualitative analysis that focuses on input restrictions for the suffix -tum. I will use all neologisms from the corpus, not just the neologisms from the 4,000,000 token subcorpora. The reason is that we would have to find a way to generalize over the potentially different sets of first attestations in the 100,000 iterations, which is not trivial. Using all neologisms with their first attestations in the larger corpus is unproblematic as long as we abstain from making quantitative statements.

34 types are first attested before 1700. If we sort them according to their respective bases, we arrive at the following set of rather diverse categories:

(2) a. Heiligtum, Eigentum, Weistum
   ‘sanctuary’, ‘wealth’, ‘property’, ‘a (wise) legal decision’

b. Altertum, Besitztum, Beweistum
   ‘antiquity’, ‘possession’, ‘evidence’

c. Herzogtum, Erzherzogtum, Burggruftum
   ‘duchy’, ‘archduchy’, ‘burggravedom’

d. Priestertum, Meistertum, Bürgermeistertum
   ‘priesthood’, ‘mastery’, ‘mayordom’

e. Christentum, Judentum, Heidentum
   ‘christendom’, ‘Jewry’, ‘heathendom’

The words in (2a) are derivations on an adjectival basis (heilig, ‘holy’, etc.); all other words have nominal bases. The bases of the words in (2b) are inanimate nouns (e.g. Besitz ‘possession’)¹⁰, the ones in (2c–2e) are animate, and more precisely, they denote persons or groups of persons. The words in (2c) are formations on titles of nobility (e.g. Herzog ‘duke’); closely related to those are the words in (2d), which have occupational titles as bases (e.g. Bürgermeister ‘mayor’). The bases in (2e), on the other hand, denote followers of belief systems (e.g. Christ ‘Christian’). Only two words do not fit well into those categories: Menschentum ‘humankind’ has as its base the most generic term for humans, while Luthertum ‘Lutherdom’ has a very specific base, namely one person, Martin Luther. All in all, the set of words used before 1700 is rather limited, but its input restrictions are very heterogeneous (adjectival and inanimate as well as animate nominal bases).

In the first half of the 18th century, only ten new types are attested. All but one (Martertum ‘affliction’) have bases that denote person nouns. In most words, this base is a title of nobility (e.g. Markgrafentum, ‘markgravedom’), or an occupational title (e.g. Großmeistertum, ‘grandmasterdom’). The base of one word denotes a follower of a belief system (Mohammedanertum ‘Islam’). The bases of the two remaining words are ethnonyms (e.g. Türkentum ‘Turkdom’). This is a new type of formation.

The second half of the 18th century also sees an expansion of input restrictions. While most of the 18 newly attested formations have bases that are either titles of nobility (e.g.

¹⁰ There are two words that would synchronically be analyzed as derivations from a verbal base, Wachstum ‘growth’ on the verbal stem wachs- ‘to grow’ and Irrtum ‘error’ on the verbal stem irr- ‘to err’. Diachronically, these words seem to be derived from nominal bases (cf. e.g. Tschentscher 1956).
Königtum ‘kingdom’ or occupational titles (e.g. Studententum ‘studenthood’), there are three exceptions to this, and they all involve an evaluation of the person in question, either positive (Heldentum ‘heroism’), or negative (e.g. Pfaffentum ‘priesthood (pejorative)’).

The input restrictions are again expanded in the next half-century 1800–1850. During this time, there are 70 new words. Bases that denote titles of nobility or occupational titles are attested, but they are in the minority (e.g. Junkertum ‘junkerdom’ or Beamtentum ‘officialdom’). The evaluative pattern is also attested with seven new words (e.g. Virtuosentum ‘virtuosity’), as is the pattern with followers of (political) belief systems (Jakobinertum ‘Jacobinism’). The largest subpattern with 15 types are ethnonyms (e.g. Römertum ‘Romanhood (i.e. Roman times)’, Persertum ‘the Persian people’). There are also a number of formations with inanimate bases (e.g. Schrifttum ‘the set of all writing (i.e. literature)’, Kirchentum ‘churchdom’) or adjectival bases (e.g. Deutschtum ‘Germanhood’). In this period, three new types of base are attested. The most prominent among them are bases that denote social status or class (e.g. Bürgerentum ‘bourgeoisie’ or Vasallentum ‘vasalldom’). A second group consists of (nominal) bases that refer to characteristic features of the person other than their origin, social status or occupation, e.g. Jungfrauentum ‘virgindom’. And finally, there is also one type with a base that refers to an animal (Bienentum ‘beedom’).

In the latter half of the 19th century, there are 241 new words; this is clearly the most productive phase of the pattern in the DTA corpus (cf. figure 6). Among the formations, we find the categories introduced above: Some bases are occupational titles (e.g. Unternehmertum, ‘entrepreneurship’); some of the bases denote a social class (Arbeitertum, ‘the collectivity of all workers’); others are ethnonyms (Japanertum, Europäertum); some are evaluative (Idiotentum, ‘idiocy’); others refer to characteristic traits of the person, e.g. Waisentum ‘orphandom’; some are titles of nobility (Zarentum ‘tsardom’), others evaluate their referent (Maulheldentum ‘the quality of being a loud-mouth’); some denote followers of belief systems (e.g. Republikanertum ‘republicanism’), yet others are inanimate (Phrasentum, Episodentum), and some denote animals (Löwentum ‘lionhood’). But again, the set of possible bases expands in this time period. There are four bases that are personal names (e.g. Wagnertum ‘Wagnerism’), and there is one base that is a cardinal number (von Humboldt’s Fünftum ‘fiveness’).

Summing up, the number of input restrictions decreases over time, and the pattern expands to other types of bases. Table 1 summarizes the findings sketched above (shaded cells indicate that a pattern is attested):12

---

11 This is not a corpus artefact – the word, even though it appears rather basic today, was indeed coined as late as the second half of the eighteenth century, cf. Tschentscher (1956).

12 The category “cardinal number” is omitted because it consists of just one type.
In the case of -tum, both \( P_{\text{neo}} \) (figure 6) and the number of types (figure 8) increase in the 19th century. This increase is mirrored by an expansion of possible bases in the same time period. Why is this interesting? Because it does not have to be this way. As long as the initial categories are large enough, an increase in productivity can theoretically happen within these initial categories. Yet it is the case here, and it would be interesting to know how often it is the case for other patterns that expand (or shrink). Once enough quantitative and qualitative analyses of the kind presented here have been conducted, we can start drawing generalizations which in turn can inform our theoretical models of productivity.

Qualitative analyses need not stop here; we could also investigate the linking element or the semantic function of -tum. There has obviously been a high degree of onomastic demand in 19th century Germany for those very abstract terms; it would be interesting to link this morphological finding to cultural studies of the time.

6. Evaluation: \( P_{\text{neo}} \) vs. \( \mathcal{P} \)

Let us take a moment to evaluate the productivity measure introduced here. The advantage of \( P_{\text{neo}} \) is that it can be interpreted as the renewal rate of the morphological category. It identifies how big a fraction of the vocabulary is new in any given time span. As such, it is a more direct measure compared with Baayen’s \( \mathcal{P} \). \( \mathcal{P} \) is defined as the quotient of the number of hapaxes with a given word formation pattern and the total number of tokens with the pattern \( \mathcal{P} = \frac{n_{1}}{N} \). The idea behind this measure is that productive patterns produce many low frequency types; the higher the amount of particular low-frequency types, namely hapax legomena, in relation to all tokens of the pattern, the higher its productivity. As mentioned above, this is an elegant solution to the problem that the newness of words is not a directly accessible category in synchronic corpora: \( \mathcal{P} \) is the probability that after sampling \( N \) tokens, the next token will be a hapax. However, \( \mathcal{P} \) as a statistical estimate is problematic because it is derived from a simple urn model. In this model, we sample from a fixed and unchanging
population. Yet words are not evenly distributed within and across texts (I would like to
thank Harald Baayen for this point).

What is more, this approach is unnecessarily complex when we are dealing with diachronic
data. As an analogy, consider marine biology. Using data about big fish (heavier than one
kilogram), it is possible to predict the number of species below this weight – e.g. 431 species
between 1 and 10 grams in the North Sea (Reuman et al. 2014). Now imagine we knew the
whereabouts of every fish in a confined area of the North Sea; would we still use the indirect
measure? We would of course use the actual data. This analogy is not as far off the mark as
it might seem. We have a sophisticated model that allows us to make an educated guess
about unknown properties (Baayen’s $\mathcal{P}$). But once we have more direct access to these
properties, the model is superfluous.

What we can do, though, is use the actual data to test the goodness of the assumption
behind $\mathcal{P}$. As hinted at above, the core idea is that hapaxes are indicative of neologisms. To
test this postulate, we simply go through every suffix and every decade and determine a) the
number of neologisms and b) the number of hapaxes in that decade. Figure 10 plots the
means values (black: new words; grey: hapax legomena) for each suffix:

![Graphs showing the number of neologisms and hapaxes per pattern per decade for different suffixes.]

Figure 10: The number of neologisms (black) and the number of hapaxes (grey) per pattern per decade for the
five derivational suffixes (-bar, -isch, -nis, -ieren, and tum).

There clearly is a correlation between both measures. This correlation is very close for some
patterns (e.g. -tum), and rather loose for others (e.g. -ieren). Of course, the values plotted in
figure 10 are the mean values over all the 100,000 simulations; they only allow for an
overview of the data. But in each simulation, we have a time series of values for new words
and one for hapaxes, and we can measure the correlation between both. That means we
have 100.000 correlation coefficients for each suffix. These values are plotted as histograms in figure 11:

![Histograms of correlation coefficients (Pearson’s r) between hapaxes and new words for the 100,000 simulations and the five derivational suffixes (-bar, -isch, -nis, -ieren, and tum).](image)

For -bar, -isch, and -tum, the correlation is rather strong (it is consistently above 0.7), with -tum showing the highest degree of correlation in the simulations. The correlations for -nis are lower, but most of them are still above 0.5. The pattern with the consistently weakest correlations is -ieren. For some patterns, we can conclude, the number of hapaxes correlates closely with the number of new words – for some, this is much less the case.

Yet even in cases where the correlation coefficients are consistently high, the actual values can differ considerably. Consider -tum and -isch in figure 10 above: For both patterns, hapaxes and new words correlate very closely (cf. Fig. 11), but the ratio between both values is very different. While the numbers of hapaxes and new words are roughly equal for -tum, the number of hapaxes is around twice the number of new words for -isch.

This holds for the other patterns as well. Figure 12 plots the quotient of both $\frac{\text{neologisms}}{\text{hapaxes}}$ for each of the suffixes:
Figure 12: The number of neologisms divided by the number of hapaxes per pattern per decade for the five derivational suffixes (-bar, -isch, -nis, -ieren, and -tum); light grey area: 99% confidence intervals, dark grey area: 95% confidence intervals.

The ratio of hapaxes and neologisms is remarkably stable over time for some of the patterns (e.g. -isch and -ieren). For -tum, however, and across suffixes, it varies considerably, from a maximum of 1 (-tum, 1850) to 0.13 (-ieren, 1750). If hapaxes are indicative of neologisms, they are so in a rather vague way.

Let us turn to a related but more specific aspect of the relation between hapaxes and neologisms. The basic assumption underlying P is that neologisms are more likely to be found among lower frequency words. With the data at hand, we can actually determine the rate of hapaxes among the neologisms for any given suffix and time span (Figure 13).
Most of the new words from each pattern are hapaxes. The ratio is relatively stable across patterns and time periods; it lies between 0.6 and 0.8 (the marginally productive pattern -nis is the only exception). This is in line with Baayen & Renouf’s (1996: 75) observation that “neologisms [...] typically occur among the lowest frequency types”. Thus if we are looking for new words, and we want to use frequency classes of words as a proxy, then hapaxes are the best choice (if 60% to 80% of all new words are hapaxes, there is not much room for dis legomena etc.).

From this does not follow, however, that hapaxes are a particularly good proxy for new words. There may be more or less hapaxes for each pattern that are not new but simply rare. This is the case in the DTA corpus, as figure 14 shows:
Thus, while it is true for the data used here that most of the new words are hapaxes, the reverse does not hold for all patterns and all time periods: Among the hapaxes, the ratio of new words varies quite drastically, between 0.6 for -tum in the 19th century and below 0.2 for -ieren. For this pattern, hapaxes are not very indicative of new words because most of the hapaxes (>80%) are simply rare but recurring words. Something similar holds for -isch, where the amount of new hapaxes lies below 40%, and for -bar and -nis, where it is below 60%. What is more, while the amount is stable over time for some patterns (-bar, -isch, -ieren), it varies for others (-nis and -tum). These differences over time and across patterns are an obstacle to the comparability of different $\mathcal{P}$ values.

Note that this may be an effect of the corpus size: It is possible that the rare hapaxes for -ieren in Figure 14 would not be hapaxes in a larger corpus, leading to a higher ratio of new hapaxes among all hapaxes. But a change in corpus size would probably also affect the near-perfect correlation between hapaxes and new words for -tum (in Figure 10 above). Different corpus sizes can amplify or mute certain metrics of patterns, and the precise change also depends on the characteristics of the pattern itself. This is in some respects analogous to room modes in acoustics: Depending on the size, shape and surfaces of a room, certain frequencies are amplified. To understand (and predict) the interaction, we need a good model of sound waves and the way they are reflected or absorbed by certain surfaces. Likewise, to understand the interaction between word formation patterns, their metrics, and corpus size, we need a good model. To build this model, we need to investigate the

Figure 14: The ratio of new hapaxes among all hapaxes per pattern per decade for the five derivational suffixes (-bar, -isch, -nis, -ieren, and -tum); light grey area: 99% confidence intervals, dark grey area: 95% confidence intervals.
relationship between corpus size, hapaxes, types, tokens, and new words more thoroughly.

Unfortunately, that is beyond the scope of this paper.

We can tentatively conclude that for different word formation patterns at different points in time, hapaxes vary with regard to the number of neologisms they are supposed to represent. Hapax-based measures should thus be used with caution in diachronic investigations. The $P_{neo}$ measure is directly interpretable, and it measures productivity more directly, as I have just demonstrated. It is also more independent of the corpus size, again in contrast to Baayen's $P$: After all, $P_{neo}$ is type-based, and $P$ is token-based. Any change in the size of the sample will affect the token-based measure naturally more.\(^\text{13}\)

7. Summary

In this paper, I have shown how a measure based on the notion of new words (which in turn are a central part of definitions of productivity) can be employed in investigations of diachronic productivity. I have presented a method that deals with the problem of varying sub-corpus sizes (randomly resample the sub-corpora up to a predefined size), and to the problem of old words appearing as new at the start of the corpus (take an earlier corpus and determine a point in time when almost all old words have registered).

The measures proposed in this paper are based on first occurrences of words in a large diachronic corpus. These measures are a more direct proxy to new words than hapax-based measures – they measures what we want to measure, which is not always true for hapax-based measures – and they are also more stable when sample sizes vary. The basic assumption of hapax-based measures (that hapaxes are indicative of new words) does not hold to the same degree for all patterns and all time periods; the correlation between new words in the corpus and hapaxes varies considerably.

Armed with the two measures based on new words, $V_{neo}$ and $P_{neo}$, I tracked changes in the productivity of five German suffixes. $V_{neo}$, the number of new words, can be used to compare patterns with respect to their contribution to the expansion of the vocabulary: From this perspective, $-isch$ plays a larger role than $-tum$ because it contributes more new words to the overall lexicon of the language. But if we look at $P_{neo}$, which can be regarded as the renewal rate of the category, $-tum$ is more productive than $-isch$, because relative to all types in the category, $-tum$ produces more new words. It is this latter measure that is more directly linked to our intuitive notion of productivity, I would argue.

The large diachronic corpora that have become available lately provide us with a richness of data that I have not done justice to. Of all the conceivable parameters of a pattern, I have only used the first attestation of word formation products. Accordingly, the end of this paper is more of a starting line than a finishing line.\(^\text{14}\) For example, we could take the last

\(^\text{13}\) Of course, this is not to say that $P_{neo}$ is independent of the sample size: the more words we sample per decade, the higher the likelihood of finding new words. That is why the number of tokens per decade should be reported (see formula (1) above). The assumption is simply that $P_{neo}$ is more stable than $P$ if we vary the sample size.

\(^\text{14}\) And if we think bigger, we can also ask for the renewal rates of larger categories, or of the vocabulary as a whole. Where do new words come from? What relative amounts do derivation, compounding, conversion, loan words, acronyms, etc. have? A prerequisite for this kind of investigation is a manually cleaned corpus, though, which is almost impossible considering the corpus size necessary for the investigation.
attestation into account and reconstruct the life-span of word-formation products (cf.

We could also think of a more sophisticated model of what it means for a word to be ‘alive’.
We could investigate the speed, or the burstiness, with which new words spread through
texts. Although not immediately relevant to productivity, this is connected to an interesting
question: What makes a new word successful? What are the (structural, pragmatic, social)
conditions for it to become lexicalized and used extensively (cf. e.g. Kerremans et al. 2012)?

References
Anderson, Karen Elizabeth (2000). Productivity in English nominal and adjectival derivation,
Anshen, Frank & Mark Aronoff (1989). Morphological productivity, word frequency and the
Oxford English Dictionary. In Fasold, Ralph. W. & Deborah Schiffrin (eds), Language
and Language 68, p. 16–26
Proceedings of the 13th international congress of linguists, p. 163–171
Publishers, p. 109–149.
Dordrecht.
Lüdeling, Anke & Merja Kytö (eds.) Corpus Linguistics: An international handbook.
Berlin: de Gruyter, p. 900–919
Baayen, R. Harald & Rochelle Lieber (1991). Productivity and English derivation: a corpus-
Baroni, Marco & Stefan Evert (2005). Testing the extrapolation quality of word frequency
models. In Danielsson, Pernilla & Martijn Wagenmakers (eds.), Proceedings of Corpus
Linguistics 2005 (http://www.birmingham.ac.uk/research/activity/corpus/publica-
Neuhochdeutschen. Submitted to Beiträge zur Geschichte der deutschen Sprache und
Literatur.


