ZW JW Zeitschrift für Wortbildung
Journal of Word Formation

2024, 8(2), 28–51
DOI: 10.21248/zwjw.2024.2.114

Hagen Peukert

Lexical Affix Productivity in the History of English: A Quantitative Approach*

Abstract: This paper addresses the development of lexical affixation throughout the last 700 years of the English language. More specifically, it pursues two objectives. First, a short outline of the methodological approaches will be devised reaching from stand-alone applications (Peukert 2014) and shared-work solutions (Peukert 2018) to requesting the OED RESTful API. Second, two sets of results will be presented. The first set includes overall aggregations of all productive affixes as well as their shares on the total number of each affix type. The second set of results elaborates on two interesting cases chosen from highly productive prefixes and suffixes. The contribution closes with a short discussion on alternative explanations and limitations of the chosen approach. Although the affix token frequencies by and large replicate the findings in Peukert (2016), which are based on type frequencies, the presented data substantiate the idea that, in terms of lexical morpheme usage, English reveals more and more characteristics of a prefixing language.

Keywords: affix productivity, diachronic analysis, derivational morphology

1. Introduction

Collecting representational quantitative data on the frequency of lexical affixes throughout 700 years of English language use has proven to be a challenging task (Dietz 2015: 1915–1917). While type frequencies of suffixes and prefixes can be determined with relative ease, the identification of token frequencies from larger text corpora employs profound computational knowledge and intensive, cumbersome methodological work. Extracting all representations of one affix type and its exact quantities requires considering all kinds of variability in form and usage. As opposed to mere type frequencies, the token frequencies are needed to make the more interesting statements on affix productivity and interrelations with other factors of influence in the system of language, i.e. a correlation to word order or predictions of likely future changes (Stein 1970; Kastovsky 2009).

* I would like to thank the two anonymous reviewers for their very constructive, usable, and concrete feedback on the first draft of this paper.

The motivation for a systematic diachronic study of affixation in English is the present state of missing data in this field. A short, but by no means exhaustive, survey of the literature reveals diachronic studies on single (productive) morphemes such as *-hood*, *-dom*, *-ship*, *-ment*, or *-age* (Ciszek 2008; Trips 2009), *-nesse* and *-ity* (Riddle 1985), *-ity* and *-ness* for Modern English (Arndt-Lappe 2014) and aggregations up to seven prefixes (Hiltunen 1983), prefixed verbs (Lutz 1997) or several suffixes (Haselow 2011). Although a vast plethora of thorough studies have been carried out, reliable statements that hold the test of representativity are rare to non-existent. To be precise, investigating a specific set of suffixes or prefixes of the past is without question valuable scientific inputs in the direction of the development of the English derivational system; yet the significance is limited for the missing context of the quantities of all other affixes. Depending on the definition of bound morphemes and whether Greek and Latin items are included, there are about 300 known affix types documented in the *OED*. This number more than triples if variation in forms is considered. Estimations on polysemous items could not yet be made, but even without those, the case is clear that statements made from very few affixes to the general behavior to all English affixes must be relativized. In other words, it is crucial to have the productivity of one affix set in relation to all other productive and non-productive affixes to really understand the underlying mechanisms.

With the above argument in mind, it follows that the main subject of the paper at hand is methodological. Hence, at first, a short survey of the attempts made and the main learnings from their failures will be given. Second, the presentation of results that proposedly come a large step closer to the ideal of a representative and somewhat contextualized data collection of English affixation. Before going to these details, background and problem space are briefly delineated. The paper closes with an abbreviated discussion of the results.

2. Problem and Background

The challenges in morphological analysis presented here are generally agreed upon (Faiß 1992; Štekauer 2000; Schmid 2016). These phenomena are replicated in all standard textbooks of this matter or encyclopedias such as Crystal (2019), so that no further

reference is made unless other information is provided. This short reproduction is provided here to be better able to relate the analytical problems to the design of the applied methods described hereafter.

The first salient characteristic of English diachronic text analysis is the high degree of variability especially in the Middle English period up to the establishment of standards by recognized dictionaries such as Johnson's *Dictionary* of 1755 (Vera 2002; Crystal 2019: 78) in Early Modern English. From this time on English writing is more homogeneous, and hence morphological analysis becomes easier. Despite missing standards in the Old English period as well, the comparatively little amount of written text and the overall conforming effect of monasteries pushes the challenge of managing text variability to the background. It remains the foremost problem of Middle English texts. Written variability mainly arises from two sources: regional differences or dialect and individual inconsistencies among scribes or even of one and the same scribe. Some scribes are known to change writing rules and styles within short time intervals. Others happen to write as they speak, and this may had led to more fluctuations within much shorter time intervals and without any observable systematic patterns of change.

Indeed, language diversity underlies known processes of language change that must especially apply in an area of extreme immigration at that time. Well-studied linguistic assimilation processes from borrowings have certainly contributed to balancing the perceived variability by the speakers. Harmonizing foreign to familiar (morphological) forms is suggested to be a psychological conformity (Ellis 2022) whereas phonetic and then phonological assimilation is due to learned physiological restriction (Blevins 2004; Antoniou et al. 2015). However, both types of assimilation may interact with each other. To illustrate, the still very productive *-er* suffix and its variant *-or* could be detected in words like *editor*. Yet, careful diachronic investigation strongly suggests that *editor* entered the English lexicon at a time in which the verb *to edit* did not exist (*OED* s.v. *edit, v.*). Hence, *editor* is not created by affixation, that is, *-or* is added to *edit*, but it needs to be identified as a backformation – a process characterized by reversed analogy to the affixation process. The form of a suffix happens to coincide with the same phonetic sequence of the ending of a lexical item, which is thus recognized as such, and accepting the remaining root, *edit*, as a new lexical entry in the lexicon.

At the same time, assimilation processes make morphological analysis more challenging. There is nothing but the very manuscript study which reveals knowledge of past assimilation. It cannot be derived from one source alone. Simply by looking at the word *establish* and many other verbs ending in *-ish* (*OED* s.v. *-ish* suffix[2]), for example, the unknowing analyst may be inclined to identify the phonetic *-ish*-sequence as a suffix. Indeed, the word was borrowed from Old French *establiss* as the lengthened stem of *establir* and was incorporated into Middle English morphology as *establisse-n* as the *OED* suggests (*OED* s.v. *establish, v.*). To assume an affixation process for Modern English is still beside the point since *-ish,* derived from Latin *-isc-*, soon became unproductive. The meaning of the Old English homonymic form, however, which transfers a noun to a corresponding adjective, kept its productivity. Both forms could be confused if changes over time remained unconsidered. In other cases, affixes may fuse with roots, stems, or other affixes. The *be*-prefix in *behead* exemplifies such a case of amalgamation. Similar to Middle High German *behoubeten*, the Old English verb *behēafdian* was once formed by prefixing the noun *hēafod*, which meant 'head' (*OED* s.v. *behead, v.*). Today the *be*-prefix became unproductive, but we still find the remainder in words like *behave* and *behavior* (*OED* s.vv. *behave, v.*; *behaviour, behavior, n.*).

The examples above illustrate the major methodological challenge for morphological analysis in general and for computational approaches in particular. Exclusive manual examination will not reach representative data unless huge amounts of human resources and time is granted. As an alternative, semi-automatic and fully automated approaches exist. In addition and because of its immense popularity nowadays, methods of Machine Learning (ML) and Artificial Intelligence (AI) are proposed for all kinds of data analysis. It is still an open question if more recent AI-technologies can be applied to a reliable identification of derivational morphemes of Middle English. Having trained an experimental supervised model, accuracy measures turned out to be low, probably due to the verb morphology. Attempts towards creating a reasonable unsupervised learning model failed as well so that these approaches are postponed to future follow-up studies.

Besides the number of word tokens in the existing corpus material, a pressing problem of computational morphology is that existing word models that define the hierarchical relations between root, stem, base, and affixes (Selkirk 1982; Booij 2010) are not

implemented as text annotations in established corpora as it is the case for the annotations of sentences (Bauer 2019: 58). Based on a solid theory such as Head-Driven Phrase Structure Grammar (Pollard & Sag 1994) or Dependency Grammar (Hays 1964), text corpora are syntactically parsed and as such can be evaluated with ease. This is not the case on the word level with no exception for diachronic corpora. In fact, rule-based approaches opt out for this very reason. While simple search algorithms collapse in very few cases on the syntactic level where annotations exist, they completely fail on the word level for the limited power of linear expressiveness. Searches mostly expressed as regular expressions are likely either to overgeneralize or undergeneralize a given population, i.e. they happen to match more words or fewer words containing the letter sequence. Since the productivity measure depends on hapax legomena, even one mishit already may lead to seriously skewed results. The following word pair in examples (1) and (2) spells out the core problem.

(1)	a.	*distemperaunce*		[inclemency]	
	b.	dis	temp	er	aunce
	c.	[dis]	[temp]	[er]	[ance]
	d.	prefix	root	stem/suffix	suffix

The French borrowing *distemperaunce*, which can be translated with ‘inclemency’ today, might be segmented as shown in (1). Seemingly, the same structure prevails in *disseveraunce* (‘separation’). It turns out that any matching algorithm using simple analogies would fail as in (2e.) through (2g.) given the variability in writing of the *dis*-prefix and the *-ance* suffix. The correct segmentation is then a matter of equally distributed probabilities.

(2)	a.	*disseveraunce*		[separation]		
	b.	dis	sever			aunce
	c.	[dis]	[sever]			[ance]
	d.	prefix	root/stem			suffix
	e.*	dis	sev	er		aunce
	f.*	dis	sever		a	unce
	g.*	diss	ev	er	a	unce

Hence, the direction of a possible solution points towards handling the diversity of affixes and the problem of embeddedness. The embeddedness problem describes the inability to recognize that a potential affix is embedded in another sequence, that is in examples (1) and (2), *-er* is used as a suffix in *temper*, but not in *sever*. Embeddedness typically occurs in replacement procedures based on regular expressions. An additional source of information could be the word class, which may change if an affix is stripped. Affixes typically are added to certain word classes but not to others. The consequences of incorporating word class information in a rule-based algorithm is twofold. First, it adds substantial complexity and, second, it reduces faultiness. There is certainly a tradeoff between these two. Complexity increases because for each word class the set of possible affixes and order information must be defined. For most prefixes and a few suffixes these sets are not disjunct.

3. Method

3.1. Methodological Assumptions and Morphological Productivity

The methodological assumptions of the study at hand hark back to determining type frequencies of affixes (Peukert 2016) but are extended with a measure of productivity. As a short wrap-up, the first assumption is that the prevalently used corpora of diachronic analysis of English (PPCME2, PPCEME, PPCMBE) are correctly parsed and are representative for the English language at that time. This assumption is strong and there are reported cases, though anecdotal, which argue against the representativeness of text corpora for diachronic analysis. This is reasonable if considering the distribution of text registers and genres in which medieval texts were written. Yet, text corpora as representatives of the language in use are the only existing source. There is little choice as to trust the engaged linguist when compiling the corpora to the rules of corpus design as best as possible (Biber 1993).

As a second assumption, the *Oxford English Dictionary* (*OED*) is acknowledged as a standard, i.e. ambiguities are resolved by consulting the lexical entry in question. However, this does not apply to word occurrences dated earlier in the corpus than claimed in the *OED*. The function of a text corpus is to balance the correct relation of actual language use as exact as possible. Technically derivational affixes are a substring of the word, so that the

frequency of the words, in which the affix occurs, is equal to the frequency of the occurring affixes. Hence, the function of the corpus (but not from the dictionary) is to provide the word frequencies, from which the affix frequencies can be calculated, by adding up all word frequencies the affix occurs in. Since this is done for all words, the quantitative relations for all affixes to each other can also be derived. Additionally, for diachronic derivational morphology, word frequencies from one period can be correlated to the frequencies of the next period. Although dictionaries also provide frequency data, the decisive difference is that the frequency data in dictionary collections is not balanced (Biber 1993).

The third assumption asserts that the assigned time slots in the corpus design do not significantly distort word frequency data. The decision of number and length of time slots, in which texts are categorized, is somewhat arbitrary a matter of agreement and plausibility. In fact, due to nonavailability of eligible texts, the amount of textual material measured in word tokens is not equally distributed among the agreed time slots. In other words, the probability of occurrence of a certain affix changes in due proportion to the size of the corpus in the respective time interval. A general and cross-linguistic property of text (Zipf 1935) is that type frequencies scale down significantly while token frequencies keep on rising relative to the text size. Since the productivity measure employed here, (3) depends on the number of tokens in the denominator; the resulting productivity values will be smaller for large texts assumed that the probability of occurrence of hapaxes – needed in the numerator – is equally likely on a normalized basis, e.g., per 10,000 words. When comparing productivity values from different text sizes, i.e. different time periods, large distances of productivity values could be treated as implied in (3), but small distances should be construed on a logarithmic scale as implied in Zipf's law relative to the text size.

(3) $$P = \frac{n_1^{aff}}{N^{aff}} \qquad 0 < \mathrm{P} \leq 1$$

For reasons of comparability and simplicity, the productivity will be defined as in (3): the number of hapaxes containing a particular affix over all token occurrences with that affix in a given text (Bauer 2001; Plag 2006; Baayen 2009: 902) and at a defined time interval.

The values range between zero and one. The closer a productivity value approaches one, the higher the productivity of the affix. The closer the value comes to zero, the lower the productivity will be. A value of zero means there is no productivity at all, which may also happen if no hapaxes with the affix are available at that specific time interval although the affix occurs frequently.

3.2. Genesis of Computational Approaches

Roughly, the work on diachronic computational morphology approaches to be described here can be summarized in three stages evolving over the last decade and revealing a constant development towards more and more degrees of automation while adopting the important insights for further improvement to the next stage. In what follows, a brief description of these stages will be provided with some more emphasis on the first stage, which is the basis of all subsequent versions.

The first stage was inspired by an old classic: the division of labor between the machine and human mind to efficiently identifying and counting affixes of some millions of words. Word parsers that give a reliable hierarchical representation of historical lexical items are up to the present effectively not available so that the best way of receiving the desired data was to have the machine do all routine work. The scientist is then free to dedicate more time to the careful analysis of the structure of words. The result was a stand-alone application called the *Morphilo* Toolset (Peukert 2012; 2014; 2016). It consisted of three components that fulfilled the machine's task of extracting the relevant lexical items (*MorExtractor*), making reasonable suggestions on the structure of these items as well as counting all representatives in the corpus at a given time spot (*Morphilizer*) and, finally, querying the data (*MorQuery*). The *Morphilizer* component contained an overgeneralizing algorithm that is still part of the current 3rd version as a robust backup in case a word could not be retrieved via the *OED* interface. Since this algorithm works astonishingly well for rare and therefore regularly formed words (Haspelmath 2008; Haspelmath & Karjus 2017: 1218–1219), it will be presented here in more detail. It should be explicitly noted that the algorithm will fail if the root of the word also happens to be a suffix or prefix form (see example (4)).

The basic idea is to approach a given word from both ends, front and back, and cut short all matches of strings from lists of prefix and suffix allomorphs after one another. The segmentation is likely to be correct if the direction of matching, i.e. start with matching the prefixes or start with matching the suffixes, has no effect on the remaining root. The root's length must also be longer than one character.

The pseudo code of the algorithm is given in abbreviated form by means of the sample word *disenablement* in (4) and Fig. 1. It is one of the rare cases, in which the algorithm does not fully succeed. It is selected here as an accessible example to the workings of the algorithm.

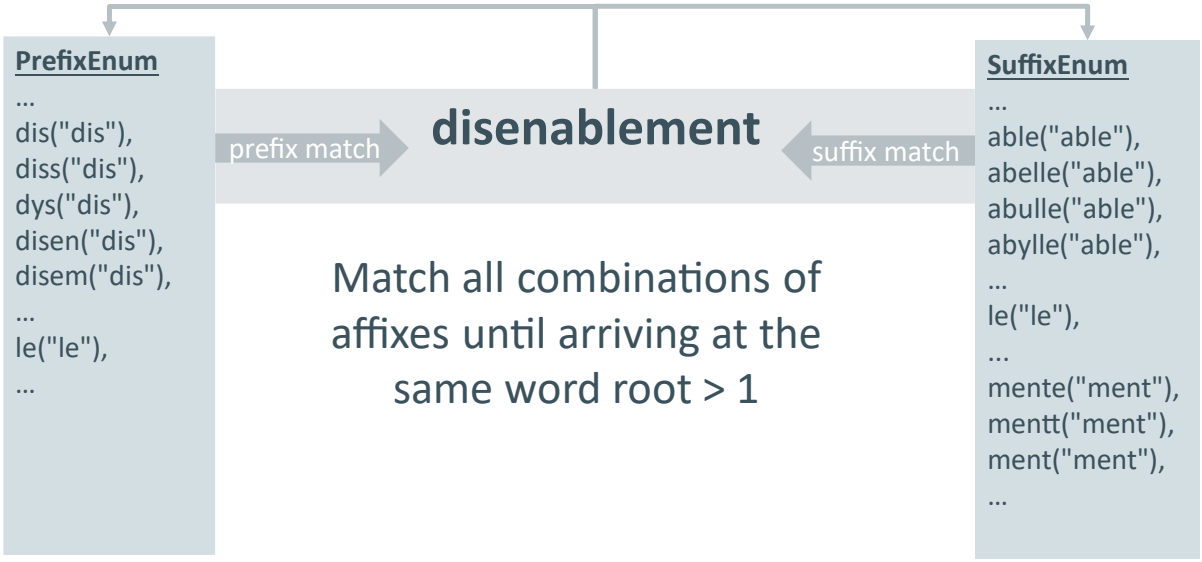

Fig. 1: Rule-based affix matching algorithm

The algorithm has access to enumerated lists of prefixes and suffixes extracted from the *OED*. These lists encode the variants of each morpheme in the program's specific syntax as *allomorph("morpheme")* pairs and can be quite long, e.g., 89 variants for the *over*-prefix. For each given morpheme, the allomorphemic variants are ordered by length. As illustrated in Fig. 1, the algorithm starts with the prefix match of the longest possible item from left to right and continues the matching process until no more matches can be made. The same will be done for the suffixes but the matching goes from right to left. Again, it is important to start with the longest match since the probability is higher for longer affixes as occurrence frequency decreases with length and so is a hit of a longer affix more likely to be correct as several shorter affixes. If the remainder of both matching processes is greater than one, the algorithm starts over in reversed order, i.e. the suffixes are firstly aligned. If the root of both

matching processes were equal, the tagged word would be kept as a likely candidate. Otherwise, the suggested segmentation will be deleted. Finally, the algorithm starts from the beginning until the enumerated lists are empty. In case of several remaining candidates at the end, the one with the longest root is presented or, as a second criterium, the number of identified affixes will be taken. The first generation of software displayed this candidate in the *Morphilizer* component for manual correction or confirmation.

(4a.) shows the correct segmentation for the word *disenablement*. In (b.) all prefixes are matched evoking the matching of suffixes in (c.). Technically the algorithm could stop here because it is already clear that there will be no roots left or they cannot be possibly equal as stated in (g.). It is revealed in (d.) through (f.), where the suffixes are matched first followed by the prefixes. The algorithm will delete the combination of affixes and repeats the above process with the next candidate from the prefix list (h.) – (j.).[1] This time, the loop ends with no roots because the form of the root happens to be a suffix as well. It also means that the algorithm in this simple form described here will never find the correct segmentation. It can only be fixed by implementing additional rules such as a preliminary checkup of monomorphemic words with the list candidates.

(4)	a.	*disenablement*	$[dis]_{pref}[en]_{pref}[able]_{root}[ment]_{suf}$
	b.	$[disen]_{pref}$ $[ab]_{pref}$ lement	no more prefixes left that match
	c.	$[disen]_{pref}$ $[ab]_{pref}$ $[le]_{suf}$ $[ment]_{suf}$	match all suffix strings
	d.	disenable$[ment]_{suf}$	reverse direction of matching
	e.	disen$[able]_{suf}$ $[ment]_{suf}$	no more suffixes left that match
	f.	$[disen]_{pref}$ $[able]_{suf}$ $[ment]_{suf}$	match all prefixes
	g.	c. and f. are not equal; no roots greater one, delete this candidate	
	h.	$[dis]_{pref}$ enablement	take next allomorph
	i.	$[dis]_{pref}$ $[en]_{pref}[able]_{suf}$ $[ment]_{suf}$	no more matches possible
	j.	$[dis]_{pref}$ $[en]_{pref}[able]_{suf}$ $[ment]_{suf}$	reverse direction of matching
	k.	i. and j. are equal; but no roots greater one, delete this candidate	

[1] The more practical implementation will start each repetition with the respective other affix, here the suffix, to encounter efficiency problems of different list sizes of prefixes and suffixes. For reasons of clarity and conciseness, it is neglected here.

Despite the semi-automated process, it became soon clear that the immense workload of analyzing word structures could still not be handled in due time by a single analyst. Consequently, the next generation of software (2nd version) needed further efficiency gains in the analysis of word structures at low costs. A promising solution at that time seemed to be a community-based approach, which would acknowledge the need of representative data in the field of historical word-formation and, at the same time, delegate some of the responsibility to each user benefiting from the data. Put briefly, other than a web-based wiki, in which each user profits from the collected knowledge without necessarily contributing to it, the new version was supposed to restrict access to active users, that is, a take-and-share approach. The result was a software called *Morphilog* (Peukert 2018)[2] that allowed all users to register via a web interface and upload part-of-speech tagged text corpora. All words in these collections would then be matched with the existing analyzed data and only those words that are missing in the master data base would be given for analysis to the user. Since Zipf's law (Zipf 1935) applies for all larger texts, the resulting set of words still to be analyzed happened to be considerably low. Once the user had completed the analysis of the missing types, the entire collection with all analyzed words including his or her own annotations would be returned. Thus, each user would only contribute a minimum of annotation work and benefit immensely from the return of the entire material. By and large a savings of 90 percent of the work could be noticed.

The architecture of *Morphilog* incorporated most of the algorithms of the first software generation but made them accessible via a web interface and an additional component that managed the quality control of newly made annotations. For the latter, a statistical solution was implemented that collected all annotations made by registered users, compared them, and wrote them to the master database if a definable limit of equal annotations was made. This limit turned out to be decisive. From an initial value of 20 equally annotated words by different users, the value was soon set to five. And even this number was rarely met. It depended crucially on the number of active users; otherwise, statistical quality control misses its point. At the end, the size of the community working in a fairly particular field and willing to trust an unknown software with questionable sustainability was the reason

[2] https://gitlab.rrz.uni-hamburg.de/mycore_projects/morphilo2019.git; https://morphilo.readthedocs.io/en/latest/index.html.

to abandon the community-based approach and return to the very roots of the project, but not without substantially reconsidering the strategy.

The availability of a RESTful API by the Online *OED* as well as a successful application for its unrestricted access initiated the starting point for yet another software version. The new and most recent software is named *Morphóchron.*[3] The procedure here is as simple as requesting data from the *OED* for each word, parsing through the result set and returning the relevant information on time slots, affixes, and roots. The fallback of items that are not listed in the *OED* reverts to the above-described *Morphilo* algorithm.

Fig. 2 depicts the architecture of software that finally generated the results that are presented in Section 4. It also serves as a description of the general procedure. The central unit is as usual the *Init* class, which after start-up invokes a graphical user interface (*GUI*). Here, the user is asked to specify the *OED* credentials, corpus, word class, and affix type. For the study at hand the *Penn-Parsed Corpora of Middle English* (PPCME2), *Early Modern English* (PPCEME), and *Modern English* (PPCMBE) as well as prefixes and suffixes and all nouns, verbs, and adjectives were selected. The *AffixStripper* class is taken from the 1st software version. All preprocessing of word classes is done with the factory design pattern (*WordClassFactory*) with a respective interface. Text normalizations are carried out in the *Corpus* class.

[3] https://gitlab.rrz.uni-hamburg.de/softwaretools/morphochron.git. The software is for public use. However, credentials for using the OED API must be separately applied for. Without the access token, *Morphóchron* will not work.

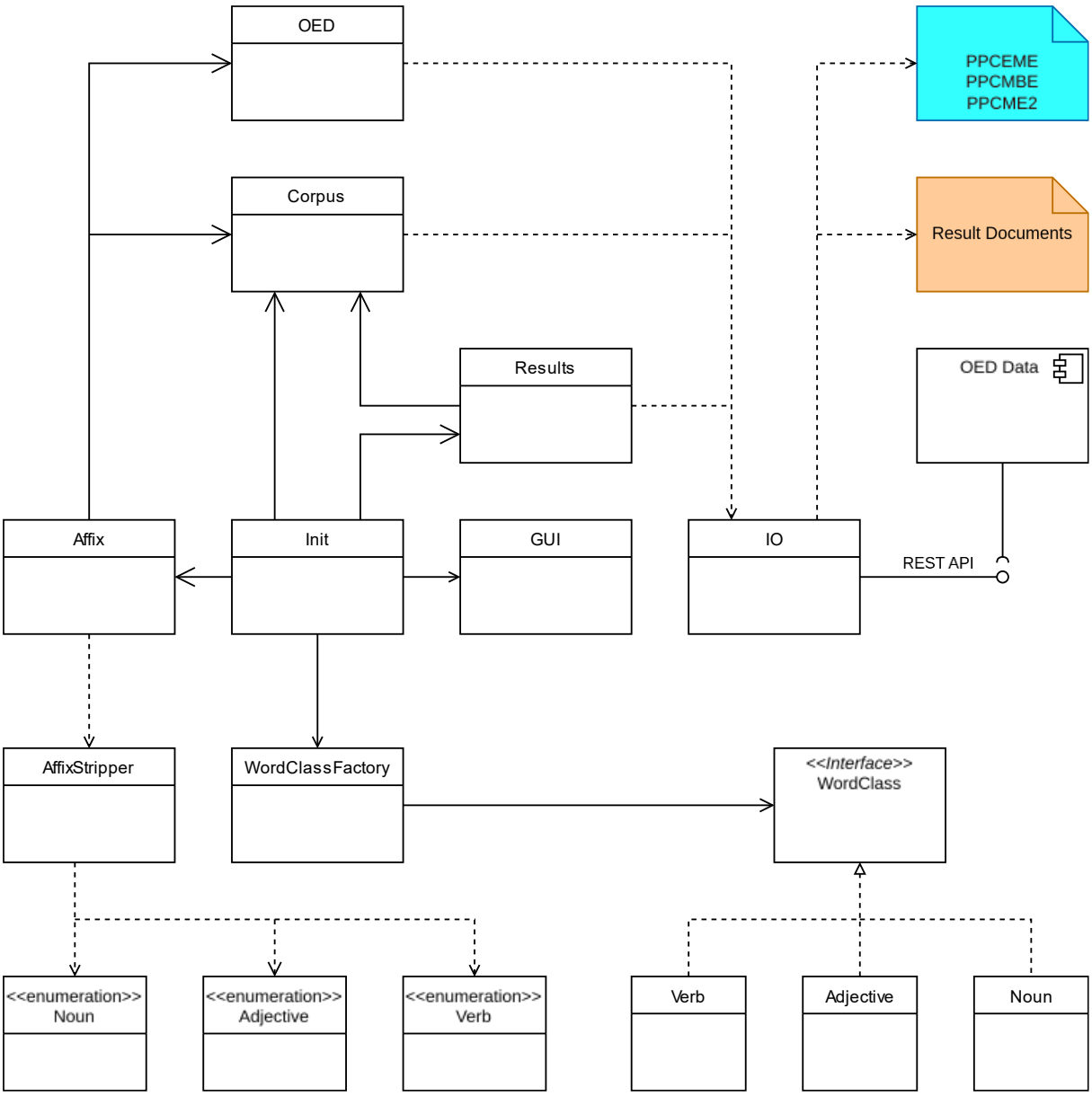

Fig. 2: Architecture of *Morphóchron*

4. Results

Morphóchron produces lists of hapaxes and words containing the respective affix.[4] For the overall analysis, the number of prefixes and suffixes of these vectors are aggregated. While affixes that only occurred once are included in the total number, they are not incorporated in the productive set although the above given definition of productivity does not prescribe that. Yet it would lead to the highest productivity value ($P = 1$) and it would distort the

[4] https://gitlab.rrz.uni-hamburg.de/softwaretools/morphochron//blob/master/Morphochron/results/resultsMorphochron.csv?ref_type=heads.

data massively. For example, the nominal suffix *-et* occurs only in the word *chapelet* once in the entire corpus (PPCME2/m3 1350–1420), which means that $n^{aff} = N^{aff} = P = 1$. In the same corpus, there are 463 nouns ending in *-ness* (159 types) from which 91 are hapax legomena. The productivity value is still below 0.2 and this would indicate a lower productivity than for *-et*. In the next period (PPCME2/m4 1420–1500), *taberette* entered the corpus as another hapax; *chapelet* accounts for two tokens, which results in a fairly high productivity measure of one third. This value comes close to the productivity of the *-ness* suffix, which occurs 122 times with 37 hapaxes. In this period *-et* is part of the productive set. While these distortions are not a problem in the present analysis, in which overall aggregations are presented and pairs of affixes are contrasted whose quantitative properties are similar, productivity classes and other measures would need to be introduced if all 379 cases were included.

Tab. 1: Absolute numbers of productive and total prefixes and suffixes

	1150–1250	**1250–1350**	**1350–1420**	**1420–1500**	**1500–1569**	**1570–1639**	**1640–1710**	**1700–1769**	**1770–1839**	**1840–1914**
prod. prefixes	22	15	18	10	5	10	15	22	31	37
prod. suffixes	30	27	42	38	25	37	40	54	77	78
total prefixes	37	37	41	35	31	42	47	62	85	106
total suffixes	43	46	64	56	52	70	87	102	124	156

If the data in Tab. 1 are sketched along the timeline (Fig. 3), one can make three important observations. First, prefixes and suffixes are on a steady rise from the 15th century on after they have gone through ups and downs in the Middle English period. As shown elsewhere (Peukert 2016) and with the possible exception of time span 1350–1420 (PPCME2/m3), the general increase cannot be explained with differing corpus sizes. Since the relation between suffixes and prefixes stemming from the same text does not dependent on the number of words, token normalization is excluded here. This leads to the second observation; the total numbers of prefixes and suffixes seem to grow by the same ratio. The development of productive suffixes and prefixes roughly follows this trend but reveals more deviation.

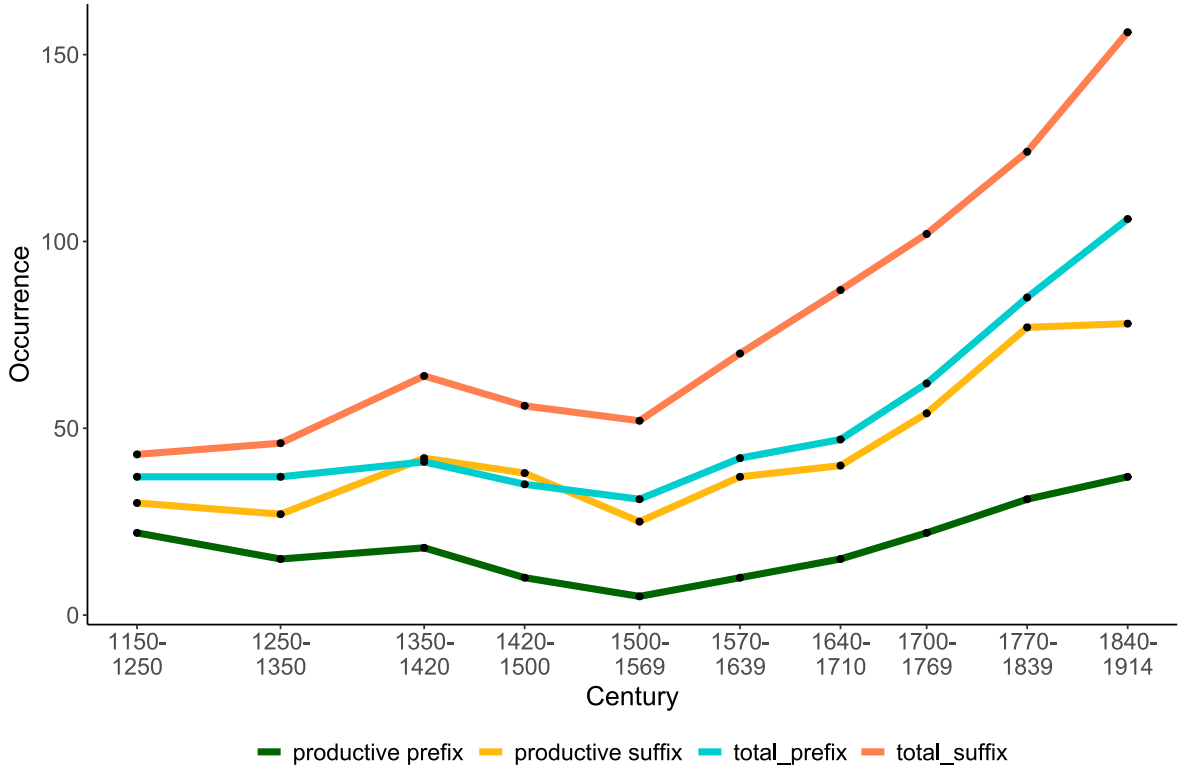

Fig. 3: Development of productive and total affixes

To understand more about the system of affixation, it is possible to relate these absolute numbers to each other. This makes the picture of what the increase means much clearer. In fact, the relation between productive suffixes and prefixes sheds light on the preferred affixation type and its gradients over time. The graphic visualization of calculating the relation between productive prefixes over productive suffixes (rel_prod), total prefixes over total suffixes (rel_tot), productive prefixes over total prefixes (rel_prod_pref), and productive suffixes over total suffixes (rel_prod_suff) is given in Fig 4.

The relation (rel_prod) shows a clear downward movement in the Middle English period. It entails that the actual use of suffixes relative to prefixes is substantially higher. After the 15th century, this tendency is reversed. There must be more productive prefixes used and created relative to the suffixes. This discovery is supported by the shares of productive affixes of all affixes (rel_prod_pref und rel_prod_suff). For prefixes, its productive share is decreasing first and then rising; the opposite is true for the productive suffix share. In this case, one could even draw a straight line at 0.5 and mirror its respective counterpart as a convex or concave function respectively. In addition, if productivity remains unconsidered and the relation of total prefixes and total suffixes (rel_tot) is

estimated, a much-flattened line with two slightly rising ends is depicted. It clearly suggests that hapax affixes did not skew the data to any larger extent. To sum up, in the last 500 years the number of (productive) suffixes grew slower than the number of (productive) prefixes.

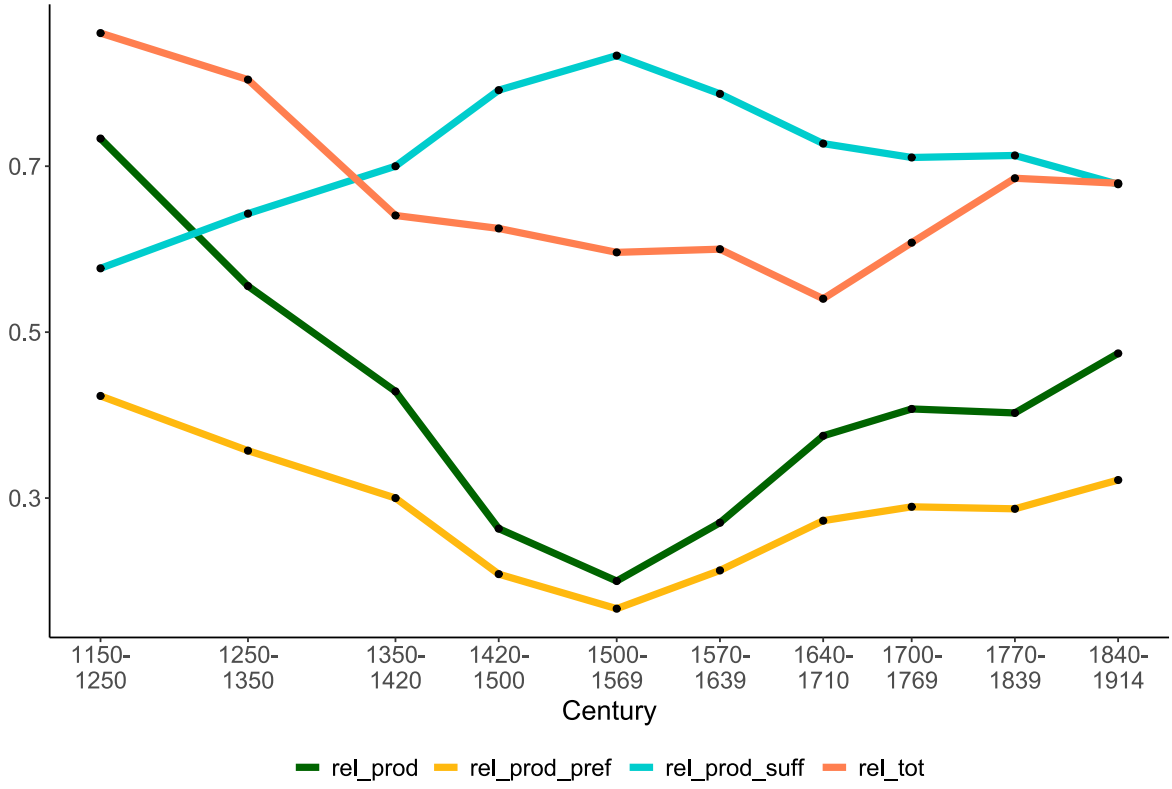

Fig. 4: Relation of productive and total affixes

The overall view generalizes from hundreds of single cases and aggregates these into a condensed picture. There is a lot of information lost on the way. Indeed, it is possible to look separately at smaller aggregates of word classes, that is, verbs, nouns, or adjectives. Moreover, it would as well be enlightening to see the effect on productivity of the affix position, syllabicity, or origin. Also, the strength of a comprehensive approach to affixation is that individual cases can be put into relation with each other.

As an illustration, the latter will be presented here. To do this, a plausible criterium should be provided. The most obvious is semantic similarity following the logic that semantically similar affixes fulfill the same function in word-formation. Substitution effects or other forms of usage behavior should then be observable. On the one hand, the prefixes *dis-* and *un-* are semantically close and at least in today's meaning distinguished

enough from other alternatives, such as *a-*, *de-*, *in-*, *non-* (Hammawand 2009: 64–72, 136). So are, on the other hand, the suffixes, *-ment* and *-ness* although to differing degrees (Schmid 2016: 169–172). It is important to note that the dependent variable is the productivity as introduced in Section 3.1. As equation (3a.) defines, the productivity score will be zero if the corpus exhibits no hapaxes, in which the affix occurs independently from the token frequency of all other words that contain the affix.

Fig. 5: Productivity scores of *dis-* and *un-* over time

The data depicted in Fig. 5 provides a first indication of a substitution effect of two negative prefixes. For about 300 years, a time of transition, in which major changes took place on various levels of Middle English on its way to Early Modern English, the usage of the Germanic prefix became unproductive. At the same time, the negative *dis-* prefix whose etymology points to Latin, gained ground on productivity by almost the same ratio. In fact, from the early 17th century on, i.e. in Modern English, the productivity of the Germanic prefix rose rather drastically to equal levels where it once started to decrease 500 years ago. At about the same time, the Latin prefix lost productivity, but was still used in new word creation processes at lower rates. It is apparent that the gradients of the two functions at the beginning (1350–1420) and the ending (1640–1770) are in inverse proportion.

Keeping the history of the British Isles and the Norman Conquest in the back of our minds (Dalton-Puffer 1996), the above data also reveals a temporally delayed shift of about 300 years until lexical affix usage is observable in text documents. According to the particular case of negative prefixes, the effect of French on the English language started suddenly but fades out long after the French influence stopped. This trend is also suggested by the suffixes although there are striking differences visible in the progression of the gradients.

Parallel to the *dis-* prefix, the *-ment* suffix enters English written material not before the beginning of the 15th century (Fig. 6). The high number of hapaxes during the next 100 years suggests highly productive usage following by an abrupt downsizing in the 16th century already and followed by an equally fading-out at lower rates as its prefix counterpart. Contrary to the *un-* prefix, the Germanic suffix *-ness* also increases productive use up to the 14th century, but completely stops being used in the 15th century before it continues at high rates of productive use for one century. With an equally steep negative slope as *-ment* one century before, *-ness* decreases and stays at lower levels of productivity.

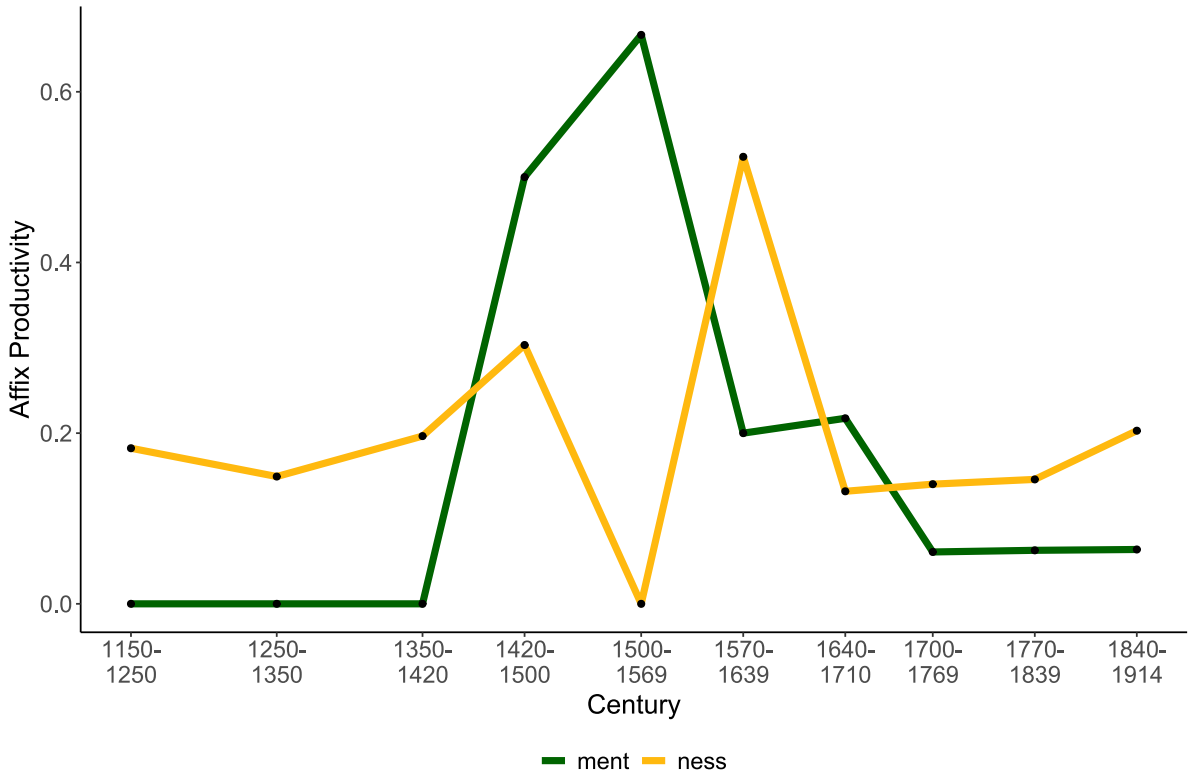

Fig. 6: Productivity scores of *-ment* and *-ness* over time

Except for the time span 1350–1420, the two suffixes show alternating productivity scores. When compared to the prefixes, it is also obvious that the declining trend of the Latin root suffix happens two centuries earlier.

5. Discussion

The above examples are a rather arbitrary selection inspired by discussions on most productive affixes encountered in the established literature. So, the purpose here was to extend this particular strain of research. Nevertheless, the *Morphóchron* data that is now available would also allow for a more systematic analysis. It would be feasible to show that no other affix is able to fulfill a likewise substitution by explicating their slopes. Further, there possibly are combined effects of several affixes substituting another affix. While these kinds of explication are left for future work, a short discussion on the presented results as well as the limitations of the approach at hand will be addressed here.

Morphóchron data does not include information on polysemy, which, arguably, could play a similar role as in lexemes (Lehrer 2003). Polysemy in derivational affixes suggests that the meaning of one affix depends on the root or base it is attached to. It also implies that this meaning can change over time for each case differently.

Considering the prefixed nouns of the last cohort 1840–1914 (PPCMBE) given as the type vectors of the algorithm for *dis-* (5a.) and for *un-* (5b.), the abstract meaning of negativity seems to hold for all items. Unfortunately, the corpora do not contain direct evidence for any cross-transfer effects of affixes with equal roots. In (6a.) *unobedience* and *unbelief* are attested. In today's dictionary *disobedience* and *disbelief* are listed (*OED* s.vv.). Hence, at some point in history a transfer occurred, which presumes semantic proximity over some constant period of time. The semantics of these cases may have been moderated by short time intervals (in the 16th and 17th centuries) of the *mis-* prefix whose semantics ('ill', 'wrong', 'improper') is often overlapping with a 'negative' prefix that switches meaning to its semantic counterpart and exists in parallel throughout the centuries with high to moderate productivity scores.

(5) a. disability, disfavour, disestablishment, dislike, distrust, discharge, disendowment, dissatisfaction, disadvantage, disorder, disappearance, discomfort, disintegration, discredit

b. unrighteousness, uneasiness, unconsciousness, unconventionality, unworthiness, unmaidenliness, unpopularity, unknowableness, unfitness, unacquaintance, unfaith, uncleanness, uncleanliness

In the 14th and 15th centuries the respective type vectors read as follows (PPCME2/m3 for *un-* and m4 for *dis-*). While *un-* reverses the meaning to its semantic opposite in all documented cases (6b.) and thus complies with its definition, it is different for *dis-* in one attested case. For *disadventure* in the reading of *misfortune* (*OED* s.v. *adventure, n.*), the definition holds. Yet, there is no reading in which the meaning of *were*, i.e. 'danger', 'peril' (*OED* s.v. *were, n.*[3]) would be directly ascribed to 'doubt' or 'hesitation' (*OED* s.v. *diswere, n.*) and there are no indicators that the affix merged with its root. Hence, *dis-* could be considered polysemous, but the problem remains which meaning *dis-* in the given sense may have instead. Whatever the correct answer to this question is, it would not make a difference for the rising productivity of *dis-* as an overall effect and with the more abstract meaning of negativity, that is, even if removing the *dis-* prefix in *diswere* as an exception or attributing it to another not yet specified meaning category would not distort the data in Fig. 5.

(6) a. disadventure, diswer

b. vnait, vnbyleue, unreste, vnknowing, vnobedience, vnreuerence

At first glance, the picture looks different for the selected suffixes *-ment* and *-ness*. It is worth mentioning that the definition of the former is narrower than the definition of the latter. Both suffixes, *-ment* and *-ness*, form abstract nouns from verbs and adjectives. However, *-ness* can also be added to participles, adjectival phrases, other nouns, pronouns, and adverbs with the consequence that in a quantitative analysis the role of robustness comes into play. The estimates for hapaxes and tokens are much higher for the established Germanic *-ness*. For example, in the 15th century (PPCME2/m4), *-ness* accounts for 37 hapaxes and 122 tokens, whereas *-ment* accounts for only four hapaxes and eight tokens. Whereas for *-ment* these numbers stay about the same in the next period, there is a

dramatic drop of *-ness* to zero hapaxes. In the 16th century then, *-ness* re-establishes to 11 hapaxes and only 21 tokens while the tokens of *-ment* rise to 30 occurrences and six hapaxes.

Left aside that the chosen productivity index does not capture robustness, the critical observations are twofold: the specific shape of the productivity's progression and the absence of usage. Suffice it here to describe these observations. The process of becoming productive is characterized by low absolute numbers of types, tokens, and hapaxes somewhere in the realm of single digits. This implies that the ratio between them and in particular hapaxes and word tokens is comparatively high. Then, after two or three centuries, the tokens rise partly exponentially, the types moderately, and the hapaxes little or not at all. Therefore, the productivity gradients increase in the initial time periods more and subsequentially flatten out. And this also applies to the investigated prefixes.

The non-usage of *-ness* in the 16th century and its revival right in the next period need more fine-grained analysis. In the case of the prefixes this period of lack of usage lasted for 400 years. Usage behavior of *-ness* seems to be more volatile than of *un-*. Apart from the 14th century, *-ment* and *-ness* exclude each other more abruptly, that is, each decrease of *-ment* is paralleled by an increase of *-ness* and vice versa. Although the observation can be explained with a substitution effect as well, it could also be a kind of phase shift in the usage of *-ness*, for which the usage of one suffix stimulates rather than substitutes the usage of the respective other. The plausibility of this argument depends critically on the explanation of the sudden lack of *-ness* usage in the 15th century. This, however, if at all, needs to be done in a follow-up study. The alternative proposals range from errors in corpus compilation over craze to historical events.

6. Summary

As laid down in Section 3, it needed several unsuccessful attempts to arrive at a method that would extract reliable data on affixes over the last 700 years from text corpora. While approaches of Artificial Intelligence and Machine Learning failed for missing sufficient training material, first semi-automated programs still needed too much manpower. Consequently, a community-based approach failed for high organizational costs and

limited ability to communicate to other researchers in the field willing to share their work and trust into an unknown resource. Finally, granted access to the OED RESTful API made the crucial difference for automating the entire extraction process and hence producing the data that would allow for answering more detailed questions in the future on how the mechanisms of derivation in English work.

Looking at productive affixes shows a general tendency. Up to the 15th century productive suffixes rose and productive prefixes declined. This process is reversed thereafter and between 1700–1914 prefixes increase faster than suffixes. Considering individual cases of frequently studied affixes, a clear substitution effect of *dis-* and *un-* is backed up with quantitative data. To some degree, the usage pattern of the prefixes is reflected in the suffixes *-ness* and *-ment*. Yet, the transition for the selected prefixes is smoother, for the suffixes more volatile.

References

Antoniou, Mark, Eric Liang, Marc Ettlinger & Patrick C. M. Wong. 2015. The Bilingual Advantage in Phonetic Learning. *Bilingualism: Language and Cognition* 18(4). 683–695. DOI: 10.1017/S1366728914000777.

Arndt-Lappe, Sabine. 2014. Analogy in Suffix Rivalry: The Case of English *-ity* and *-ness*. *English Language and Linguistics* 18(3). 497–548.

Baayen, Harald R. 2009. Corpus Linguistics in Morphology: Morphological Productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, Volume 2, 899–919. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110213881.2.

Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Bauer, Laurie. 2019. *Rethinking Morphology*. Edinburgh: Edinburgh University Press.

Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4). 243–257. DOI: 10.1093/llc/8.4.243.

Blevins, Juliette. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.

Booij, Gert. 2010. *Construction Morphology*. Oxford: Oxford University Press.

Ciszek, Ewa. 2008. *Word Derivation in Early Middle English*. Frankfurt: Lang.

Crystal, David. 2018. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Dalton-Puffer, Christiane. 1996. *The French Influence on Middle English Morphology: A Corpus-Based Study of Derivation*. New York: De Gruyter Mouton.

Dietz, Klaus. 2015. Historical Word-Formation in English. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-Formation: An International Handbook of the*

Languages of Europe, Volume 3, 1914–1930. Berlin, Munich & Boston: De Gruyter Mouton. DOI: 10.1515/9783110375732-021.

Ellis, Nick C. 2022. Second Language Learning of Morphology. *Journal of the European Second Language Association* 6(1). 34–59.

Faiß, Klaus. 1992. *English Historical Morphology and Word Formation: Loss versus Enrichment*. Trier: Wissenschaftlicher Verlag Trier.

Hamawand, Zeki. 2009. *The Semantics of English Negative Prefixes*. London: Equinox.

Haselow, Alexander. 2011. *Typological Changes in the Lexicon: Analytic Tendencies in English Noun Formation* (Topics in English Linguistics 72). Berlin & New York: De Gruyter Mouton.

Haspelmath, Martin. 2008. Frequency vs. Iconicity in Explaining Grammatical Asymmetries. *Cognitive Linguistics* 19(1). 1–33. DOI: 10.1515/COG.2008.001.

Haspelmath, Martin & Aandres Karjus. 2017. Explaining Asymmetries in Number Marking: Singulatives, Pluratives, and Usage Frequency. *Linguistics* 55(6). 1213–1235. DOI: 10.1515/ling-2017-0026.

Hays, David G. 1964. Dependency Theory: A Formalism and Some Observations. *Language* 40(4). 511–525. DOI: 10.2307/411934.

Hiltunen, Risto. 1983. *The Decline of the Prefixes and the Beginnings of the English Phrasal Verb: The Evidence from Some Old and Early Middle English Texts*. Turku: Turun Yliopisto.

Kastovsky, Dieter. 2006. Typological Changes in Derivational Morphology. In Ans van Kemenade & Bettelou Los (eds.), *The Handbook of the History of English*, 151–176. Newark: Wiley-Blackwell.

Lehrer, Adrienne. 2003. Polysemy in Derivational Affixes. In Brigitte I. Nerlich, Zazie Todd, Vimala Herman & David D. Clarke (eds.), *Polysemy: Flexible Patterns of Meaning in Mind and Language*, 217–232. Berlin & New York: De Gruyter Mouton. DOI: 10.1515/9783110895698.217.

Lutz, Angelika. 1997. Sound Change, Word Formation and the Lexicon: The History of the English Prefix Verbs. *English Studies* 78(3). 258–290.

OED = Oxford English Dictionary. https://www.oed.com (accessed December 2023).

Peukert, Hagen. 2014. The Morphilo Toolset: Handling the Diversity of English Historical Texts. In Anne Ammermann, Alexander Brock, Jana Pflaeging & Peter Schildhauer (eds.), *Facets of Linguistics* (*Language and Text Studies*), 161–172. Frankfurt: Lang.

Peukert, Hagen. 2016. Smoothing Derivational Asymmetries in English: In Support of Greenberg's Universal 27. *STUF – Language Typology and Universals* 69(4). 517–545.

Peukert, Hagen. 2018. Merging Community Knowledge and Self-Interest to Build Language Resources: Architecture and Quality Management of a Take-and-Share-Approach of Word Annotations. In Manuel Burghardt & Claudia Müller-Birn (eds.), *INF-DH-2018*. Bonn: Gesellschaft für Informatik. DOI: 10.18420/infdh2018-01.

Pollard, Carl & Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Plag, Ingo. 2006. Productivity. In Bas Aarts & April McMahon (eds.), *Handbook of English Linguistics*, 537–556. Oxford: Blackwell.

Riddle, Elizabeth M. 1985. A Historical Perspective on the Productivity of the Suffixes *-ness* and *-ity*. In Jacek Fisiak (ed.), *Historical Semantics – Historical Word-Formation*, 435–462. New York: De Gruyter Mouton.

Schmid, Hans-Jörg. 2016. *English Morphology and Word-Formation: An Introduction*. Berlin: Schmidt.

Selkirk, Elizabeth O. 1982. *The Syntax of Words* (Linguistic Inquiry Monographs 7). Cambridge, MA: MIT Press.

Stein, Gabriele. 1970. Zur Typologie der Suffixentstehung (Französisch, Englisch, Deutsch). *Indogermanische Forschungen* 75. 131–165.

Štekauer, Pavol. *English Word Formation: A History of Research (1960–1995)*. Tübingen: Narr.

Trips, Carola. 2009. *Lexical Semantics and Diachronic Morphology: The Development of* -hood, -dom *and* -ship *in the History of English* (Linguistische Arbeiten 527). Berlin & New York: Niemeyer.

Vera, Javiar E. Díaz (ed.). 2002. *A Changing World of Words: Studies in English Lexicography, Lexicology and Semantics*. Amsterdam & New York: Rodopi.

Zipf, George K. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Mifflin.

Hagen Peukert
Universität Hamburg
Zentrum für nachhaltiges Forschungsdatenmanagement
Monetastraße 4
20146 Hamburg, Deutschland
hagen.peukert@uni-hamburg.de