NATURAL SELECTION IN SELF-ORGANIZING MORPHOLOGICAL SYSTEMS

Mark Lindsay and Mark Aronoff
Stony Brook University

Abstract

Productive elements in a language compete with each other for productivity because of three central aspects of the language system: the introduction of random elements, the propagation of a suffix via productive derivation, and the intolerance of synonymy that can lead to the productive death of a less robust suffix. We investigate this emergent phenomenon, which parallels natural selection, in borrowed suffixes of English. We look historically at the emergence of productivity of *-ity*, *-ment*, and *-ation*, and see that *-ment* failed when fewer verbs were available, while *-ation* words were being borrowed into English in far superior numbers. Next, we examine *-ic* and *-ical* using data from Google search results to show that, while *-ic* is more productive overall, *-ical* is far more productive with stems ending in *-olog-*; this morphological niche was able to form because the *-olog-* subset is sufficiently large and has remarkably few neighbors. Finally, we explored the domains of *-ize* and *-ify* in a manner similar to *-ic* and *-ical*. Although *-ize* is preferred in a vast majority of words, *-ify* is dominant in the phonologically defined domain of monosyllabic stems.

1 Introduction

Evolution is a natural process; biological ecosystems organize themselves through the process of natural selection. Because of evolution, complex systems can arise (and change) from the sum of numerous smaller interactions. When a system has an element of random variation in its agents, and the traits of some agents allow them to persist while other less successful agents become extinct, then it follows that an emergent process similar to natural selection should guide the system. In language, the lexicon comprises such a system. Neologisms, speech errors and borrowings continually introduce random change into the system. An active affix survives productively in the lexicon by attaching itself to new words in the system; conversely, if the traits of a productive affix do not

allow it to derive new elements as effectively as other affixes, it will ultimately cease to be productive, though previously created forms may remain fossilized in the lexicon.

In this paper, we will investigate evidence for emergence within the English suffix organization, focusing primarily on borrowed suffixes: when new potential suffixes enter the system through whole-word borrowings, the system ultimately organizes these elements; as the language evolves, the organization of elements adapts to these changes. These emergent processes in the English suffix system are a part of glossogenetic evolution (Hurford 1990, also discussed in Steels 1997 and Fitch 2010, among others), a concept that is distinct from phylogenetic evolution, i.e. the evolution of the human language faculty.

1.1 Productivity

A productive morphological pattern is one "[that] can be extended to new cases, can be used to form new words" (Booij 2005). Usually, existing words in the lexicon play a key role in extending a productive pattern; this is particularly true of affixes, which are bound morphemes that cannot function as the root of a word. Continued productivity of an affix over a long period of time depends on the continued introduction of new words into the language that the affix can combine with. In addition, the frequency of existing affixed words in the language establishes the exemplars used to extend the pattern (for extended discussion, see Bauer 2001).

Thus, an affix that maximizes its pool of qualifying new words to attach to, maximizes its body of existing words containing the affix, and minimizes its restrictions (phonological, semantic, pragmatic, and so on) will maximize its ability to be generally productive in the language. Should the language change in any way that shifts these distributions, it could have an impact on the productivity and behavior of the suffix.

1.2 Competition

Competition is common among productive morphemes in a language; affixes that are synonymous can be said to be in direct competition with each other. For example, the suffixes *-ity* and *-ness* both convert adjectives into a nominal form and have many similar semantic properties. A speaker must choose one of these

suffixes; the exemplars from previous use influence this choice, and this choice itself serves to influence future use.

The primary driving force behind competition in the lexicon of a language is that, in general, languages do not tolerate true synonymy; in the case of affixes, that means that one *stem+affix* combination will be preferred over another combination. Therefore, synonymous productive affixes are competing for a limited resource: new words with which to combine. While any single competing affix can win out for a particular word, only one affix will dominate a particular domain. If the other affix does not differentiate itself from the dominant affix in some fashion, the less competitive affix is doomed to permanently lose its productivity.

1.3 Randomness

Through borrowing from other languages, the coining of new words, speech errors, or reanalysis of existing words, new forms can develop. For example, the suffix *-ic* initially entered into the English language when a large number of words ending in *-ic* or *-ique* began to be borrowed from French and Latin (Marchand 1969). The pattern established by these words ultimately led to the development of *-ic* into a productive morpheme of English. These new elements can cause instability in a system. An established productive affix could eventually be ousted from its place by whole-word borrowings from other languages or a significant change in word frequency in the relevant domain. We see evidence for this in the loss of productivity of the suffix *-ment*, discussed in detail in Section 2.

1.4 Adaptation

Although one affix will tend to dominate a broad domain, a language can settle into a stable system that includes the less competitive affixes as productive elements. This is achieved if the less-productive affix happens to find a niche: a clearly defined subdomain within its potential domain — a subsystem that is therefore distinct and predictable to a speaker in spite of a general trend towards another affix. Furthermore, in order for an affix to remain productive, this subdomain must also be a large enough subset of all eligible words that speakers can generalize its usage and that the affix will have an ongoing inflow of new words to combine with.

This subdomain can be defined along various conditions:

a. *phonological:* The usage of an affix can be restricted by stress, syllables, or prosody (of the stem or the output of the affixed word). A phonologically restricted affix does not entail a poorly suited suffix; in fact, these restrictions can serve to strengthen the affix within a limited (but sustaining) domain. For example, suffixes *-ize* and *-ify* are able to co-exist because of their restrictions (discussed in further detail in Section 4).
b. *morphological:* An affix can attach strictly to another specific affix, a phenomenon known as potentiation (discussed in further detail in Section 3).
c. *pragmatic:* An affix can be constrained to a specific register. For example, one might productively use a Latinate or French suffix such as *-esque* or *-ian* in a formal register, but use *-ish* in an informal register.
d. *semantic:* In this case the affixes actually settle into a situation in which they are no longer competing; i.e., the attachment of suffix *A* to word *X* does not preclude the attachment of suffix *B* to the same word *X*. For example, suffixes *-hood* and *-ship* both originally meant "state or condition", but *-ship* is now restricted to a "stage-level" interpretation, while *-hood* can have a stage-level or individual-level interpretation (Aronoff and Cho 2001).

1.5 Individual Words

Even if a suffix is productive, higher-frequency words using the suffix are nonetheless stored in the lexicon (Stemberger & MacWhinney 1988). We can see evidence of this in the suffix pair *-ic* and *-ical*. The words *historic* and *historical* have different meanings, as do *electric* and *electrical,* but the difference between the *-ic* and *-ical* forms cannot be generalized across these words; instead, the meanings are specific to the words themselves. Therefore, the words in these (high frequency) doublets must be individually stored in the lexicon, even though all contain productive affixes.

If two competing affixes survive in a language because one of the suffixes has found a niche for itself, then both suffixes are nonetheless "available" in the grammar in some sense. Thus, although there will be overwhelming preference for one form over another (e.g. *electronic* rather than *electronical*), in a corpus that is broad enough and large enough, it would be likely to see some uses of the

non-preferred form (e.g. 0.25% in the case of *electronical*). In many cases, a "mutation" such as this may never become lexicalized. But, occasionally, an individual word can settle into a separate space from the dominant form, by distinguishing itself semantically or pragmatically. In the case of *electric/electrical* and *historic/historical*, the words in each doublet have settled into distinct semantic domains.

2 Borrowed suffixes -ment, -ity, and -ation

This investigation (extending results from Anshen & Aronoff 1999) explores why *-ment* lost its productivity while *-ity* has survived to the present day, using data from the Oxford English Dictionary[1]. This data from the OED captures both the birth of productive affixes in a language and their divergence in productivity over a period of several hundred years. We will show that *-ment*'s decline in productivity was likely caused by the combination several factors,

Both *-ment* and *-ity* originally entered into English from whole-word French borrowings. The suffix *-ment* had an earlier start, with a significant number of borrowings beginning before the 14th century:

> "*-ment* is a substantival suffix, chiefly forming deverbal nouns from Romance roots. It came into the language through loans from continental Old French and Anglo-French." (Marchand 1969: 331)

English did not see a substantial number of borrowings of *-ity* words until later:

> "*-ity* forms abstract substantives from adjectives with the meaning 'state, quality, condition of –'… The oldest words are 14th and 15th century loans from French…" (Marchand 1969: 312)

[1] In Anshen & Aronoff (1999), the OED on CD-ROM, 2nd Edition, was used, while in this investigation, information was gathered from the OED website (*oed.com*).

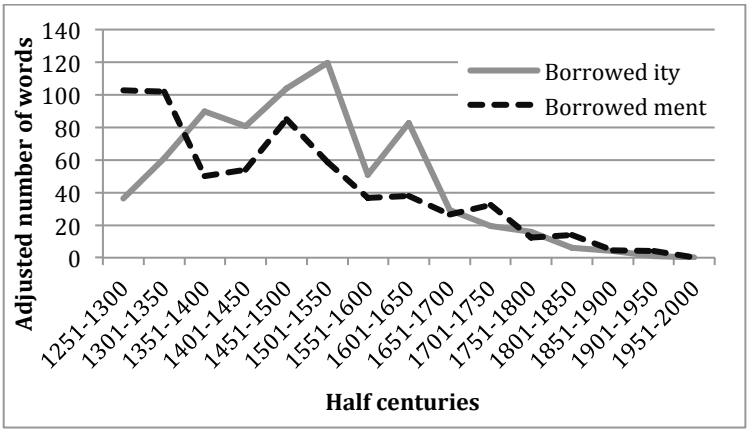

Figure 1: borrowed -ity vs borrowed -ment (adjusted[2]).

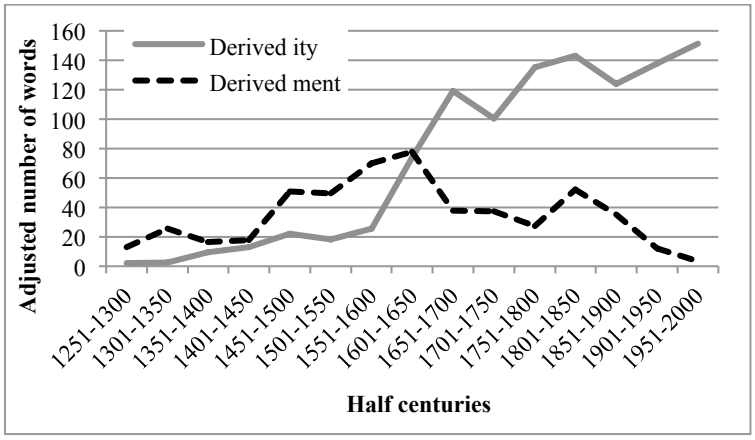

Figure 2: derived *-ity* vs. derived *-ment* (adjusted).

[2] The number of words has been adjusted to account for the varying number of words recorded in the OED over this span of time; the value for the number of words in a given half-century is proportional to the total words in the OED for that time period:

$$\text{adjusted number of words} = (\text{number of words} / \text{total words}) \times 10^5$$

In Figure 1, we see the rate of borrowings of words containing *-ment* and *-ity* from 1250 to 2000 (adjusted for the total number of words entering the OED during that time period). Both of these suffixes had a large number of borrowings from French early on, followed, unsurprisingly, by a gradual decline; the number of new borrowings reached nearly zero by the end of the 20th century.

In Figure 2, we see that the fate of these two suffixes was quite different. Early on, few words of English were derived using *-ment* or *-ity* as a productive suffix. The number of derivations increased, presumably as the number of exemplars increased as a result of continued French borrowings. However, in the early 17th century, the productivity of *-ity* and *-ment* began to change drastically. While *-ity* flourished, creating hundreds of new derived forms, *-ment* began a decline that has resulted in zero derived forms by the present day.

Why did *-ity* sustain itself as a productive affix while *-ment* failed? By their very nature, productive affixes exist in morphological ecosystems, where they depend on new words as sources for sustained productivity. These two suffixes had different productive "niches": *-ity* attached to adjectives (e.g. *equal* → *equality*) and *-ment* attached to verbs (e.g. *punish* → *punishment*). As we see in Figure 3, the number of new verbs entering English took a sharp decline in the 17th century, which is the same time that *-ment* began its decline in productivity:

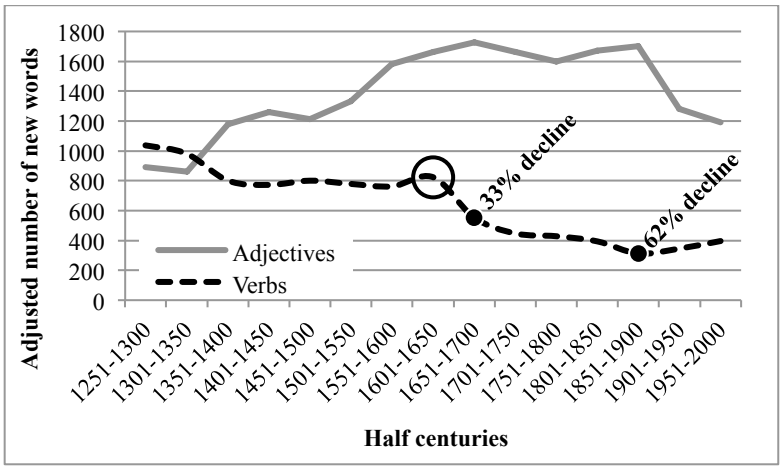

Figure 3: total new adjectives vs. new verbs per half-century (adjusted).

During the latter half of the 17th century, the number of new verbs decreased by one-third, and by the middle of the 19th century, this number had decreased by nearly two-thirds.

While this drastic reduction in new English verbs is striking, it cannot be the only factor that led to the failure of *-ment* productively; indeed, no matter how few new verbs enter a language, there remains a fundamental need for transforming verbs into nouns. In fact, a competing suffix had begun to take hold, namely, *-ation*.

> "*-ation* anglicizes [Latin] *-atio* as well as (learned) [French] *-ation*, but is now largely an independent suffix with impersonal deverbal substantives." (Marchand 1969: 258)

As we can see in Figure 4, *-ation*, like *-ity*, surpassed *-ment* during the 17th century and continued to be used to derive an increasing number of forms:

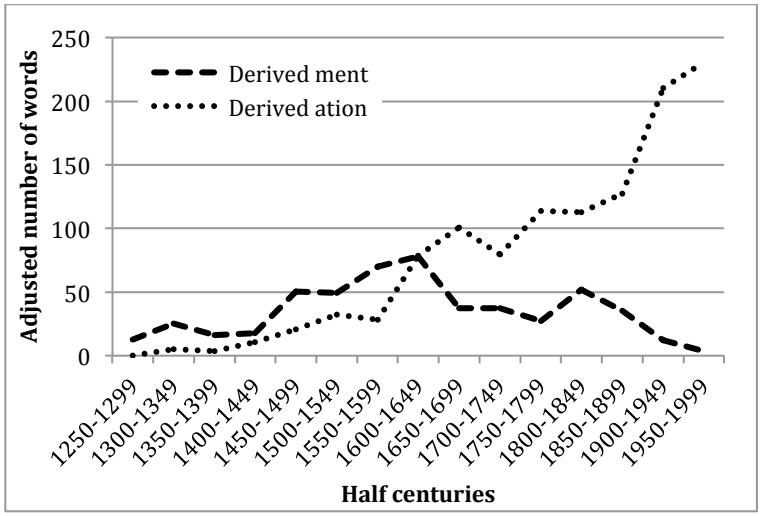

Figure 4: derived *-ation* vs. *-ment* per half-century (adjusted).

Like *-ment*, *-ation* also depends on verbs to derive new forms. However, whole-word borrowings of *-ation* words began later than for *-ment* and *-ity*, and, crucially, continued at a high rate during this critical period in which *-ment* began its decline:

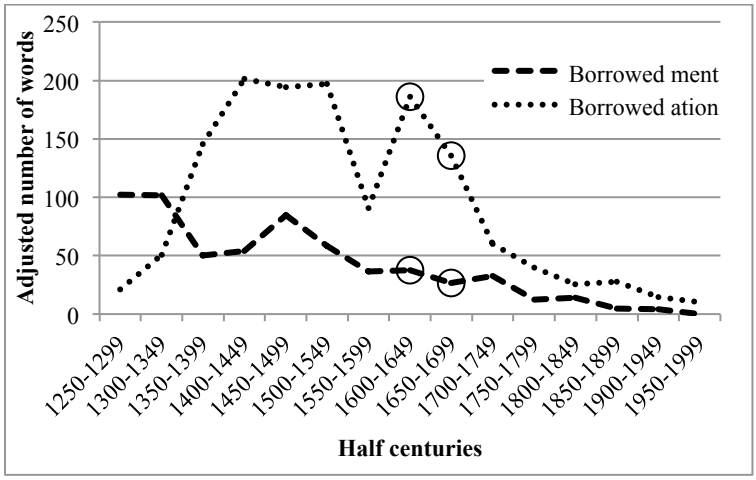

Figure 5: borrowed *-ation* vs. *-ment* per half-century (adjusted).

There are two noteworthy observations about the data in Figure 5. The first is that *-ation* words were being borrowed into English at a higher rate than *-ment* for a number of centuries. The second, and most important, observation is that, during the 17th century, there were five times as many borrowings containing *-ation* as there were containing *-ment*. This meant that *-ation* had a significantly higher level of support through borrowings when resources became scarcer for derivations, driving *-ment* towards its productive death, accelerating a process that was already in progress.

3 Morphological niche: suffixes *-ic* and *-ical*

> "There was, at the beginning, indiscriminate coexistence of two synonymous adjectives. But language does not like to have two words for one and the same notion, and competition was bound to come." (Marchand 1969: 241)

The suffix pair *-ic* and *-ical* can be considered 'rivals', because they are synonymous, and, like *-ment* and *-ation*, are in direct competition with each other for productivity. However, in contrast to *-ment* and *-ation*, both suffixes appear to be highly productive today.

In this section, we investigate the structure of the self-organizing system that has come to support both *-ic* and *-ical* as productive members. If these suffixes are truly synonymous in their basic function, then their coexistence must result from other factors.

The suffix *-ic* ultimately comes from the Greek suffix *-ikós*, but entered English via French *-ique*, which in turn borrowed the suffix from Latin *-icus* (Marchand 1969: 294).

On the other hand, *-ical* is an English creation, albeit a hybrid of two borrowed suffixes. The suffix *-al* came into English from French, though "neither the OED nor the grammars say anything convincing as to how *-al* became an English formative" (Marchand 1969: 236). Later, *-ical* was reanalyzed as a suffix in its own right due to the extensive early use of *-al* with the names of sciences ending in *-ic(s)*, such as *mathematical*, *poetical*, *geographical*, and so on (Marchand 1969: 241).[3]

In present-day usage, we find many *-ic*/*-ical* doublets (e.g. *symmetric*/*symmetrical*, *historic*/*historical*, and *electric*/*electrical*). As discussed in greater detail in Section 1.5, while the two forms in doublets like *historic*/*historical* have different meanings, the differences in these forms cannot be generalized to a difference between the suffixes *-ic* and *-ical* themselves. In most *-ic*/*-ical* pairs, one form seems to be strongly preferred over the other (e.g. *electronic* over *electronical*, *surgical* over *surgic*, and *atomic* over *atomical*).

3.1 Measuring -ic and -ical

To evaluate of each of these competing suffixes, we measure productivity in a novel way, by incorporating Google search results[4]: the exact literal string for

---

[3] Interestingly, although *-ical* was extensively used with the names of sciences ending in *-ic(s)* originally, this does not seem to be true in the present day. An analysis of *-ic*/*-ical* words that have a corresponding noun ending in *-ics* resulted in *-ic* being favored over *-ical* by a ratio of 4.3 to 1. Thus, even though *-ical* owes its creation to *-ics* nouns, the system has since evolved in a different direction.

[4] Other measures of morphological productivity exist, such as Baayen (1993), Plag (1999) and Bauer (2001). The approach used in this proposal is not meant to replace currently existing methods; rather, it is an additional means of measuring productivity that exploits the vast amount of linguistic information contained within the World Wide Web.

words is queried, and the Estimated Total Hits (ETM) result from Google is recorded for each word; we then look for numerical patterns in these numbers to determine productivity.[5]

Using basic regular expression matching, we identified all words ending in either *-ic* or *-ical* (or both) in Webster's 2nd International Dictionary and stripped off the suffixes to produce 11966 unique stems. Using the Google Search API[6], we then executed automated queries for each stem and suffix combination (e.g. *biolog-* + {*-ic*,*-ical*}) and recorded Google ETM values in a database.

In order to establish productivity measures across the entire range of data, we compared the *-ic* and *-ical* forms for each stem pair; the form with the highest ETM value was considered the "winner" of that pair.

3.2 Results

For some stems, a Google query yielded a high number of results for both suffixes (Table 1); however, for other stems, one suffix yielded far more results than the other (Table 2). In general, most stems clearly favored one suffix over the other; in fact, of the 11966 pairs, 10729 (88.5%) pairs had counts that differed by at least one order of magnitude.

---

[5] One must be cautious when incorporating Google ETM values into a measurement of usage. While Google is a vast and freely-available resource, it is also "noisy"; that is, individual results contain false positives due to typos, non-native speech, spam, the lack of part-of-speech tagging, and so on. Furthermore, ETM results represent the number of pages a string is estimated to appear in, not the number of occurrences. (Other discussion of such considerations can be found in Hathout and Tanguy 2002, among others.) For these reasons, it is important that little weight is placed upon the actual raw numbers themselves (only relative differences should be considered) or upon any individual word pairs. For the time being, it is also important to restrict investigations to single words, rather than phrases, due to the algorithm by which Google estimates phrasal results. A broad investigation of suffixes mitigates many of these concerns, as we are dealing with single words, regular inflection patterns, and a large number of stems.

[6] This research draws on data provided by the University Research Program for Google Search, a service provided by Google to promote a greater common understanding of the web.

stem	*-ic* count	*-ical* count	*ratio (-ic/-ical)*
electr-	325,000,000	218,000,000	1.49
histor-	133,000,000	258,000,000	0.52
numer-	23,900,000	37,200,000	0.64
logist-	13,000,000	5,850,000	2.22
asymmetr-	10,400,000	6,410,000	1.62
geolog-	7,980,000	22,800,000	0.35

Table 1: Sample Google ETM counts for high-frequency doublets.

stem	*-ic* count	*-ical* count	*ratio (-ic/-ical)*
civ-	90,000,000	2,220	40,540
olymp-	73,300,000	1,130	64,867
polyphon-	32,800,000	869	37,744
sulfur-	10,600,000	0	—
mathemat-	1,740,000	48,900,000	3.56×10^{-2}
typ-	421,000	158,000,000	2.66×10^{-3}
theolog-	71,300	18,100,000	3.94×10^{-3}
post-surg-	287	1,090,000	2.63×10^{-4}

Table 2: Sample Google ETM counts for high-frequency singletons.

Overall, we identified 10613 "winners" favoring -*ic* and 1353 favoring -*ical*, for an overall ratio of 7.84 in favor of -*ic*. This demonstrates conclusively that, by this measure of productivity, -*ic* is more productive than -*ical*.

If we filter out all pairs in which the winner differed from the loser by less than an order of magnitude, then the ratio tilts further in favor of -*ic* at 11.56.

3.3 Neighborhoods

Why does -*ical* appear to endure as a productive suffix today, given the fact that -*ic* is far more productive? Since language does not tolerate synonymy, such a clear preference for one suffix over the other should lead to the demise of -*ical*. Upon closer inspection, we find evidence of potentiation (Williams 1981) within the data. To investigate this, the 11966 stems were sorted into right-to-left alphabetical neighborhoods. For example, the set of all stems ending in -*t*- (neighborhood length 1) has 4166 members, while the set of all stems ending in -*graph*- (neighborhood length 5) has 294 members, and the set of stems ending in -*mat*- (neighborhood length 3) has 399 members. In Table 3, we can see the largest set at each neighborhood length:

neighborhood length	*set*	*number of members*
1	*-t-*	4166
	average	*2033*
2	*-st-*	1129
	average	*387*
3	*-ist-*	660
	average	*133*
4	***-olog-***	**475**
	average	*62*
5	*-graph-*	294
	average	*26*

Table 3: sets with the largest number of members according to neighborhood length.

3.3.1 The -olog- set

After evaluating all possible neighborhood sets in the data, only one set with a significant number of members was found to favor -*ical* over -*ic*: the -*olog*- set.

In this set, -*ical* was the winner over -*ic* by a ratio of 6.42, nearly equal to the inverse of the ratio of the full set of stems (7.84 in favor of -*ic*). Again, if we filter out pairs that differ by less than an order of magnitude, the ratio becomes 17.44 in favor of -*ical*.

Although -*olog*- is the largest set with a neighborhood length of 4, having 475 members (versus an average of 62), no other large sets have resisted the overall trend favoring -*ic* over -*ical*. However, the -*olog*- set is unique in another way: it has strikingly few neighbors. For example, as we see in the Euler diagram in Figure 6, there are 79 stems ending in -*rist*-, but 660 ending in -*ist*-; this number jumps to 4166 in the -*t*- set. This means that the -*rist*- set makes up just 1.9% of all stems ending in -*t*-.

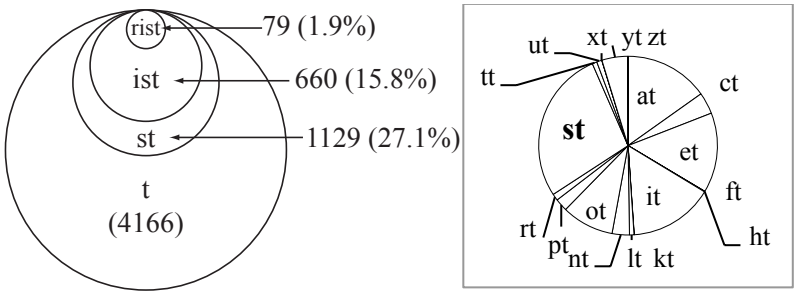

Figure 6: size of neighborhoods in a sample set, -*rist*- (left); neighbors of -*st*- in neighborhood length 2.

In the pie chart on the right in Figure 6, we can see the large number of neighbors with a neighborhood length of 2; though -*st*- is the largest subset of -*t*-, -*at*-, -*ot*-, -*it*-, and -*et*- are also large subsets, along with many other minor subsets.

On average, a set with a neighborhood of length 2 is 27.8% of its length-1 superset, as shown in Figure 7. A set with a neighborhood length of 4 is 10.5% of its length-1 superset.

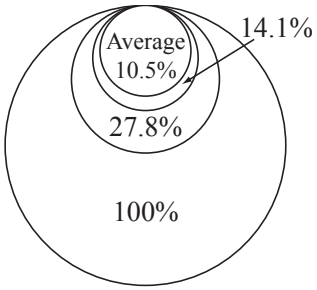

Figure 7: size of neighborhoods on average.

However, even at length 4, the -*olog*- set still makes up 66.6% of its length-1 superset, as we see in the Euler diagram in Figure 8.

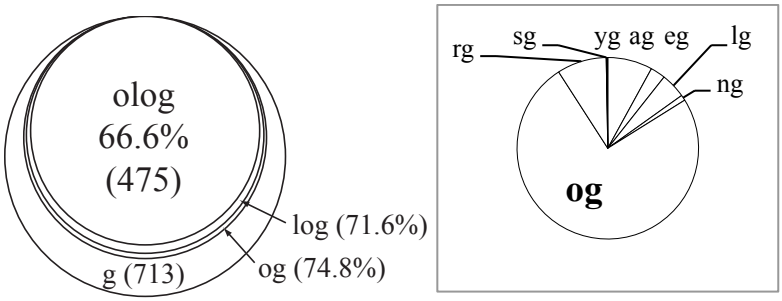

Figure 8: size of neighborhoods in -*olog*- set (left); neighbors of -*og*- in neighborhood length 2.

This means that 66.6% of all stems ending in -*g*- also end in -*olog*-, which exceeds all length-4 sets by a wide margin (the closest competitor being -*graph*- at 34%). Indeed, in the pie chart on the right in Figure 8, we see that there are few neighbors, and none that rival -*og*- in size.

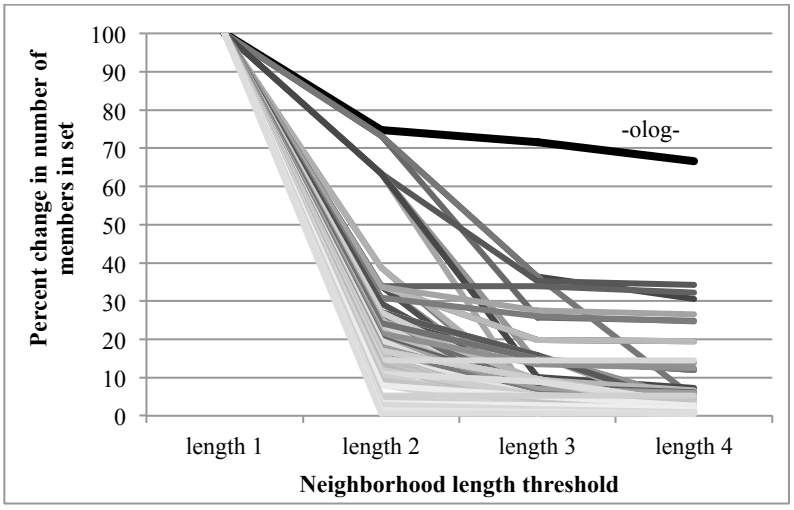

Figure 9: -*olog*- has few neighbors compared to all other sets.

Thus, the -*olog*- set is a morphologically defined subsystem that is not only sufficiently large, but also has distinctly few neighbors, leaving it uniquely suited to sustain -*ical* as a productive suffix in spite of the clear dominance of -*ic* overall.

3.3.2 Discussion

If words ending in -*ic* or -*ical* form a simple emergent system, we might predict, based on the overall prevalence of -*ic* words, that this rival would eventually win out and that -*ical* would lose. Instead, we see that a strong regularity, even one that is the reverse of the normal pattern, can develop in a subset if the subset stands out.

One might initially suspect that the reason for -*olog*- words to prefer -*ical* is the fact that the alternative is for those words to end in -*logic*; as this is already a high-frequency noun, there may be some form of blocking effect that prevents the -*ic* suffix from being preferred, leaving -*ical* as the only viable alternative. Indeed, words like *musical* would seem to support this: high frequency *music* functions as a noun and almost never as an adjective.

However, there are numerous exceptions that call such a hypothesis into serious doubt. Words like *public* and *plastic* are not only used as adjectives as well as nouns, but their -*ical* counterparts, **publical* and **plastical*, are virtually unheard of. Further, words for languages and ethnic groups are used interchangeably as nouns and adjectives, likewise with essentially no -*ical* forms: *Arabic*, *Icelandic*, *Nordic*, *Slavic*, *Semitic*. In a self-organizing system, such an explanation is not necessary to motivate the formation of a morphological niche; the initial cause may have been little more than chance.

3.4 Specialized use of -ic

While the -*ical* suffix is vastly preferred over -*ic* in words containing -*olog*-, there is one pragmatically defined domain where, at least anecdotally, -*ic* seems to be preferred.

> "…the scholar uses the unextended [-*ic*] forms much more, as for him the quality expressed by the adjective is more directly and intimately connected with the thing to which it is applied than it is for a non-scientist…" (Marchand 1969: 242)

Occasionally, the more marked form will be chosen in an academic, technical, or otherwise formal context. This may also be the case for such suffix pairs as -*ity* and -*ness*, where the -*ity* form is preferred formally, even where -*ness* is otherwise preferred. If true, these are examples of pragmatic domains (as mentioned in Section 1.4).

4 Phonological niches for suffixes -*ize* and -*ify*

The pair of -*ize* and -*ify* represents another rivalry. The -*ize* suffix originated in Greek, but both suffixes came into English via French and Latin. They convert nouns and adjectives to verbs, with the meanings "render, make, convert into" (Marchand 1969). Like, -*ic* and -*ical*, there is usually a strong preference for one or the other for a given stem. In fact, this preference may be even more pronounced for -*ize* and -*ify*, with vastly fewer doublets than the previous suffix pair.

In this investigation, the methods of Section 3 were repeated, comparing the number of Google Estimated Total Matches for each form in a rival pair. In order

to compare stem syllable systematically, two guidelines needed to be followed in order to make a fair and consistent assessment:

1. In many cases, the stem differs in structure (prosodic structure, segments) between the suffixed form and the standalone word (should one exist). For example, disyllabic words such as *simple* and *deity* have suffixed forms *simpl-ify* and *de-ify*. In this case, the suffixed form of the stem was used for syllable count; therefore, *simplify* and *deify* were considered to have monosyllabic stems.

2. Order of affixation was also taken into account. Verbs like *simplify* and *stabilize* can take prefixes to create forms such as *oversimplify* and *destabilize*. As the meanings of *oversimplify* and *destabilize* clearly show their connection to *simplify* and *stabilize*, and as the words *oversimple* and *destable* do not exist, the forms *oversimplify* and *destabilize* should be categorized as a monosyllabic stem and disyllabic stem, respectively.

Out of 2636 unique stems ending in either -*ize* or -*ify*, 2217 favored -*ize* in head-to-head competition, while 419 favored -*ify*, yielding an approximate 5:1 ratio in favor of -*ize*.

However, these results are governed by phonological restrictions. If these results are reorganized according to the number of syllables in the stem, a different pattern emerges. We find that, while polysyllabic stems still favor -*ize* (2127 -*ize* vs. 89 -*ify*), monosyllabic stems overwhelmingly favor -*ify*. In this subset, there were 322 -*ify* winners versus only 68 -*ize* winners; thus, by a ratio of nearly 5:1, -*ify* is favored over -*ize* in this domain.[7]

[7] In an investigation presented by Lignon (this volume) on French suffixes -*iser* and -*ifier*, the same tendency, though weaker, was found in French. Suffix -*iser* was preferred approximately 90% of the time with polysyllabic stems, while -*ifier* was preferred 55% of the time (compare to 82% of the time in English). Assuming that this tendency was already in place when English began borrowing these words into the language, then the preponderance of polysyllabic stems in -*iser* words and monosyllabic stems in -*ifier* words might have provided the initial template by which English then organized its suffixes. And, not only did English organize in the same way, it actually strengthened these domains beyond French itself.

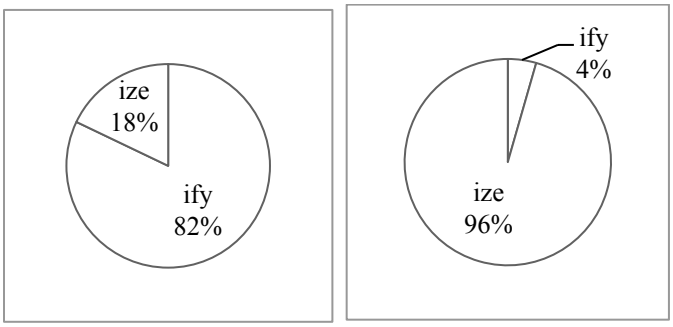

Figure 10: *-ize*/*-ify* winners for monosyllabic stems (left) and polysyllabic stems (right).

If we look more closely at the number of syllables in the stem, we can see that there is not simply a dichotomy between monosyllabic and polysyllabic stems. Rather, as the number of stem syllables increases, the tendency towards *-ify* drops off logarithmically, as we see below:

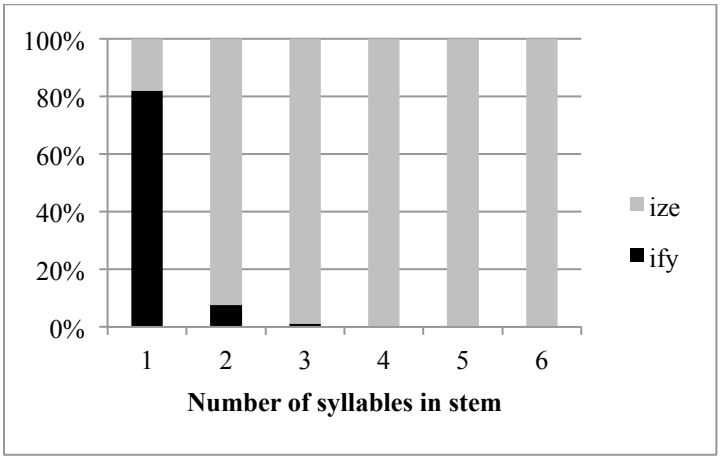

Figure 11: distribution of *-ize* and *-ify* "winners" by number of syllables in stem.

As with *-ic* and *-ical*, we have a large and clearly-defined subset of all possible words. While *-ic* and *-ical*'s subsets were defined along morphological

boundaries, the -*ize* and -*ify* subsets are constrained phonologically. Each achieves the same result: two competing suffixes are able to coexist in a system that does not tolerate synonymous affixes by finding a niche within a subset of their domain.

5 Conclusion

Productive elements in a language compete with each other for productivity. This results from three central aspects of the language system: the continuous introduction of random elements, the propagation of a suffix via productive derivation, and the intolerance of synonymy that can lead to the productive death of a less robust suffix.

First, with -*ment*, -*ity*, and -*ation*, we saw the emergence of productivity out of whole-word borrowings from French and Latin. Though each suffix began to derive new English words, only -*ity* and -*ation* survived to the present day. The failure of -*ment* was due, in part, to the sharp decrease in new verbs entering English, as well as the simultaneous borrowing of rival -*ation* words at a much higher rate. It became impossible for -*ment* to compete, and its productivity ultimately ceased.

In the case of -*ic* and -*ical*, we saw that rival suffixes can coexist, even if one suffix is clearly more productive overall. While -*ic* is clearly preferred in general, -*ical* survives productively because it has carved out a morphologically defined niche, namely, stems ending in -*olog*-. This subset is not only significantly large, but also has strikingly few neighbors. Productivity was measured using data from comparisons of relative differences in Google Estimated Total Matches (ETM) over a large number of words.

The -*ize* and -*ify* suffix pair evolved a niche similarly to -*ic* and -*ical*, but in this case the less generally productive -*ify* established a phonologically defined domain in monosyllabic stems. As with -*ic* and -*ical*, this organization arose out of nothing more than whole-word borrowings containing these suffixes.

Overall, and somewhat surprisingly, English derivational morphology, especially when it involves the emergence of productive affixes from sets of borrowed words (in which English is especially rich), is a fertile proving ground for the study of self-organizing systems in languages, in part because of the databases that electronic resources provide.

In further investigations, we hope to explore other English suffix rivalries, such as *-ity* and *-ness* (e.g. *readability* and *happiness*); we predict similar circumstances surrounding their co-existence. We will also look cross-linguistically at suffixes that have diverged in related languages. The triplet *-dom*, *-hood*, and *-ship* has counterparts in German (*-tum*, *-heit*, and *-schaft*), Dutch (*-dom*, *-heid*, *-schap*), and other Germanic languages. The role of each of these three suffixes is unclear; are they truly rivals? By comparing their distributions in English, German, and Dutch, we hope to determine the domain and function of these suffixes, and the extent to which each differs from its counterparts in its sister languages. Ultimately, this approach may be able to inform the problem of suffix ordering restrictions (analyzed recently in Plag and Baayen 2009).

We also intend to compare these results to data in traditional corpora, such as The Corpus of Contemporary American English (COCA) and The Corpus of Historical American English (COHA), as a means of verifying and supplementing the results of this investigation.

6 References

Anshen, Frank, Mark Aronoff. 1999. "Using dictionaries to study the mental lexicon". *Brain and Language* 68.10.

Aronoff, Mark and Sungeun Cho. 2001. "The Semantics of -ship Suffixation". *Linguistic Inquiry* 32.1: 167-173.

Baayen, R. Harald. 1993. "On frequency, transparency, and productivity" *in* Booij, G. and J. van Marle (eds.), *Yearbook of morphology 1992*. Dordrecht: Kluwer. 181-208.

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

Booij, Geert. 2005. *The Grammar of Words: An Introduction to Linguistic Morphology*. Oxford: Oxford University Press.Plag 1999

Fitch, W. Tecumseh. 2010. *The Evolution of Language.* New York: Cambridge University Press.

Hathout, N., L. Tanguy. 2002. "Webaffix: finding and validating morphological links on the WWW". *Proceedings of the Third International Conference on*

Language Resources and Evaluation. Las Palmas de Gran Canaria, Espagne, 1799-1804.

Hurford, J. 1990. "Nativist and functional explanations in language acquisition" *in* I. M. Roca (ed.), *Logical Issues in Language Acquisition*. Dordrecht: Foris Publications, 85-136.

Lignon, Stéphanie. This volume. "-iser and -ifier suffixations in French: verify data to verize hypotheses".

Marchand, Hans. 1969. *The categories and types of present-day English word-formation. A synchronic-diachronic approach*. Munich, Germany: Beck.

Plag, I., R. Harald Baayen. 2009. "Suffix ordering and morphological processing". *Language*, 85: 106- 149.

Plag, Ingo (1999). *Structural constraints in English derivation*. Berlin: Mouton de Gruyter.

Steels, L. (1997) "The Synthetic Modeling of Language Origins" *in* Gouzoules, H. (ed), *Evolution of Communication, vol. 1, nr. 1*. Amsterdam: John Benjamins Publishing Company, 1-34.

Stemberger, Joseph Paul and MacWhinney. 1988. "Are Inflected Forms Stored in the Lexicon?" *in* M. Hammond, & M. Noonan (ed.), *Theoretical Morphology*. New York: Academic Press, 101-116.

Williams, Edwin. 1981. "Argument structure and Morphology". *Linguistic Review* 1: 81-114.