Natural selection in self-organizing morphological systems

Mark Lindsay
Stony Brook University

Mark Aronoff
Stony Brook University

In recent years, it has been shown that certain subsystems of human languages can be profitably investigated in terms of self-organizing or emergent systems (Baayen (1993), Anshen and Aronoff (1999), Albright (2002), Plag and Baayen (2009)). Using new methods made available by the Internet and modern databases, we will show that the birth of productive affixes from borrowed vocabulary can be treated as an emergent system, where affixes survive or perish due to circumstances of environment and competition. We present two types of evidence for this position. The first is historical and addresses the question of how a language borrows affixes at all. We examine the history of two sample suffixes in English, *-ment* and *-ity*, both of which came into English from French, using the OED as a database. The second type of evidence is a synchronic investigation of the rival suffix pair *-ic* and *-ical*, using Google as a measurement of relative productivity.

Borrowed Suffixes *-ment* and *-ity*

In Figure 1, we see the rate of borrowings of words containing *-ment* and *-ity* from 1250 to the present day (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; this decline reached nearly zero by the end of the 20th century.

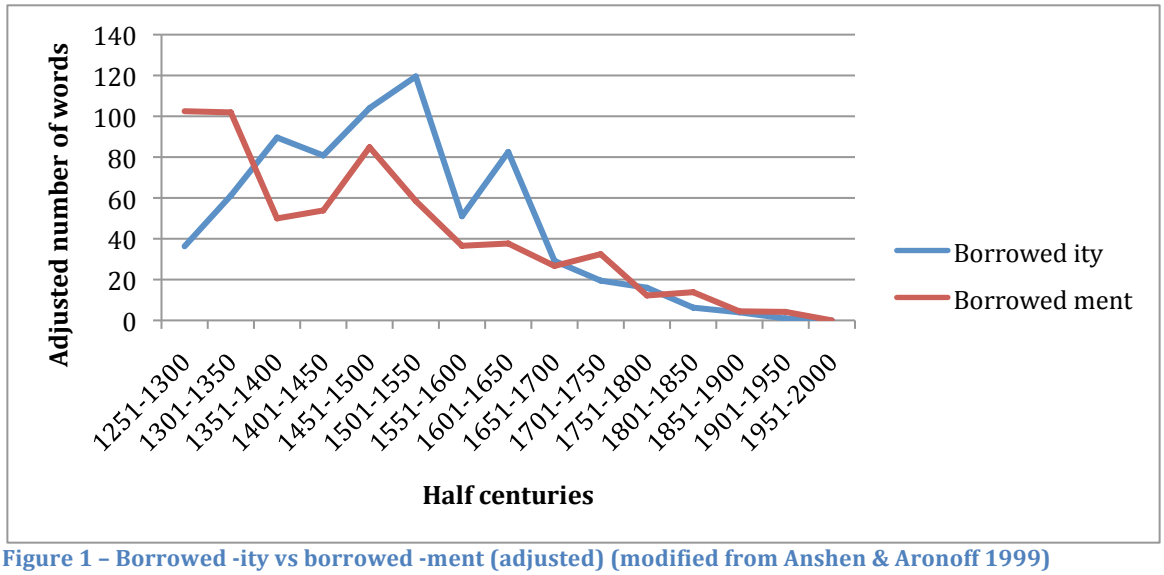

Figure 1 – Borrowed -ity vs borrowed -ment (adjusted) (modified from Anshen & Aronoff 1999)

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.

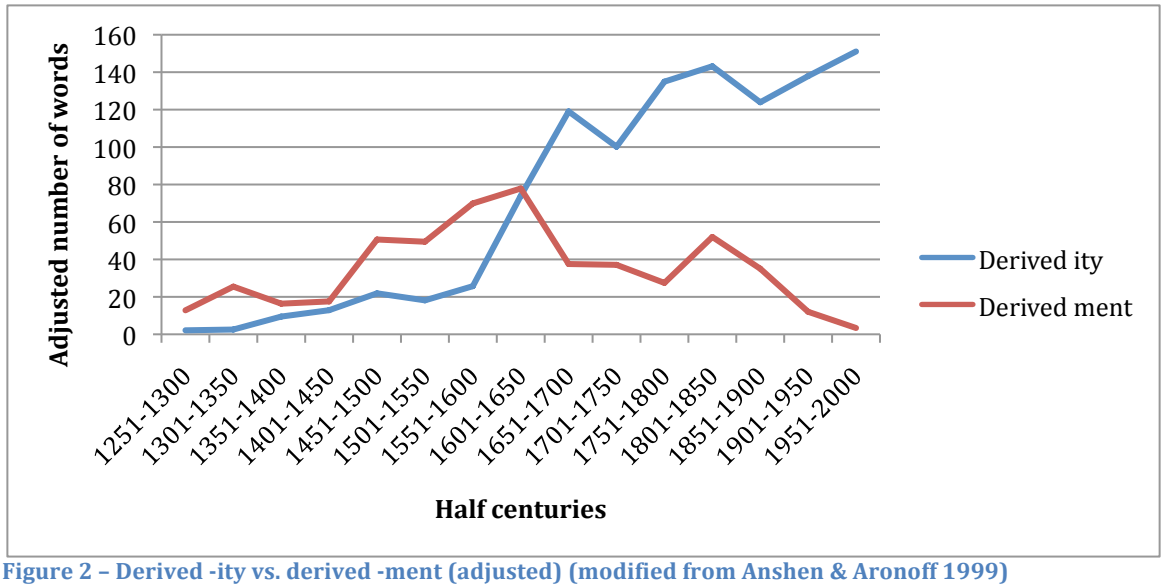

Figure 2 – Derived -ity vs. derived -ment (adjusted) (modified from Anshen & Aronoff 1999)

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 continued 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*). We see in Figure 3 that the number of new verbs declined over the centuries, with its most significant dip taking place in the 17th century, the same time that *-ment* began its decline in productivity. On the other hand, the number of adjectives increased and remained far above the number of verbs. Thus, *-ment* died out because it could no longer be sustained in its ecosystem.

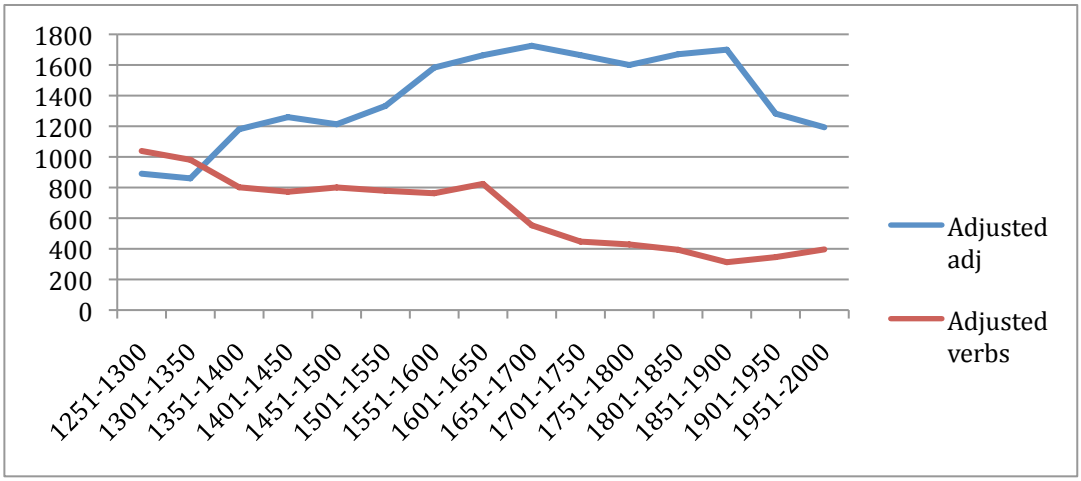

Figure 3 – Total new adjectives vs. new verbs per half-century (adjusted)

Rival suffixes *-ic* and *-ical*

Another type of evidence comes from the study of the rival suffix pair *-ic* and *-ical*. We call the pair 'rivals' because they are synonymous, and thus are in direct competition with each other. Why would a language tolerate synonymous affixes? The question is even more mysterious in this case because, although both have their origins in borrowed suffixes (Greek *–ik* and Latin *–al*), *-ical* is more or less an English creation.

Although the language exhibits doublets, e.g. *historic* and *historical*, the distinctions in meaning or usage exist only in the full word pairs; the difference between *historic* and *historical* cannot be generalized to predict that of, e.g., *electric* and *electrical*.

Furthermore, in most pairs, one form is strongly preferred over the other: *electronic* is much more common than *electronical*, while *surgical* is much more common than *surgic*. The first question we ask is whether there are grounds for saying that one suffix is more productive than the other. What is novel about this research is the way in which we measure productivity: the number of Google hits for each word ending in each suffix. For every word, we do a Google search on the exact term and list the number of hits; we then look for numerical patterns in these numbers to determine productivity.

Using basic regular expression matching, we identified a total of 11966 stems of English words ending in either *-ic* or *-ical* in Webster's Second International dictionary. For each stem, either one or both derivatives was listed in the dictionary. We then ran automated Google queries for words ending in both *-ic* and *-ical* in each of the 11966 stems, storing the results in a database. We then determined for each pair whether *-ic* or *-ical* had more hits; this word was considered the 'winner'. For most stems, a Google search resulted in hits for both *-ic* and *-ical* words (e.g. *historic* and *historical*), while for a relatively small number, only one word had any hits (e.g. *radiosurgical/*radiosurgic* but *overenthusiastic/*overenthusiastical*). Ninety percent of all comparisons yielded a winner by a margin of at least one order of magnitude. Overall, we identified 10613 *-ic* winners vs. 1353 *-ical* winners with an overall ratio of 7.84 in favor of *-ic*. This demonstrates conclusively that, overall, *-ic* is more productive than *-ical*.

Finer-grained analysis, however, reveals a subtler story — namely that *-ical* is potentiating (Williams 1981) within a certain domain. To investigate this, we first sorted all the stems in our database into left-to-right alphabetical neighborhoods of one to five letters, e.g. all stems ending in *-t-* (there are 4166 of these), or all stems ending in *–graph-* (there are 294 of these). We find that, when we sort the words in this way, the only set of words ending in *-ical* with a neighborhood over 100 in size is *–olog(ical)*; for this subset only, *-ical* is the winner over *-ic* (e.g. *psychological* over *psychologic*) by a ratio of 8.30, almost the exact reverse of the ratio of the full set (7.84 in favor of *-ic*). In other words, using this Google search technique, we find that, although overall *-ic* is more productive than *-ical*, the reverse is true for words ending in *-olog(ical)*.

Although *-olog(ical)* forms a large subset (475 members), no other large sets in the system have resisted the overall trend favoring *-ic*. However, the *-olog(ical)* subset is unique in another way: the string 'olog' has a strikingly low number of competitors. For example, there are 79 stems ending in *-rist-*, but 660 stems ending in the substring *–ist-*, and this number jumps to 4166 stem ending in *–t-*. Thus, *-rist-* makes up only 1.9% of stems ending in *-t-*, leaving many similar competitors. We found that, on average, a neighborhood of length 2 (e.g. *-st-*) only accounts for 27.84% of the words in the corresponding neighborhood of length 1 (e.g. *-t-*), while a neighborhood of length 4 makes up just 10.47% of its size-2 set. However, even at length 4, the *-olog-* subset still makes up 66.62% of its length 1 (*-g-*) set. This means that 66% of words ending in *-gic(al)* also end in *-ologic(al)*, which is the strongest of all sets by a wide margin (the next strongest is *-graph-* at 34%).

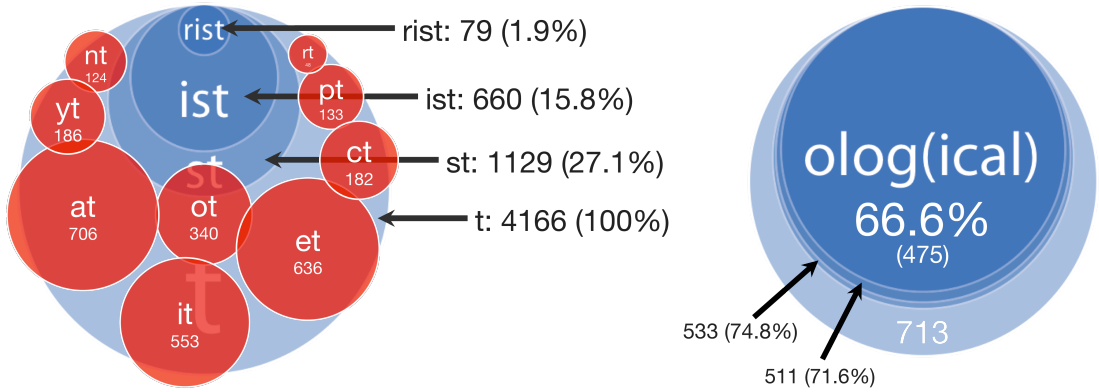

Figure 4 – Competing similar forms in the -rist- neighhorhood vs. the relative uniformity of -olog-

Thus, *-olog(ical)* is a subsystem that is not only sufficiently large, but also has distinctly few competitors, leaving it uniquely suited to sustain itself in spite of patterning inversely to all other *-ic/ical* pairs.

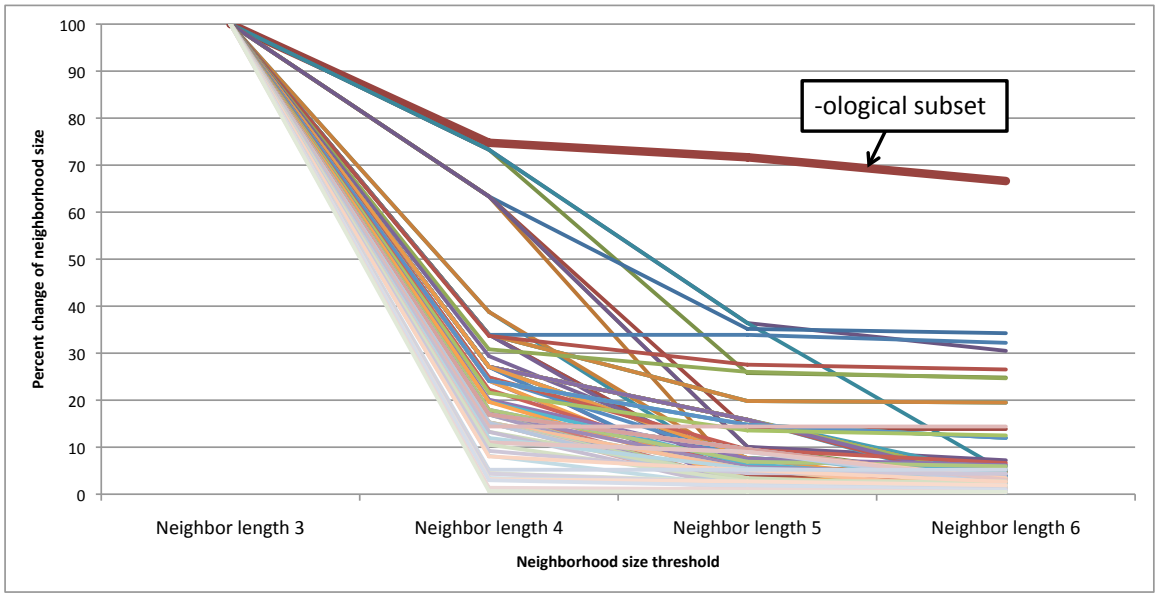

Figure 5 – Change in neighborhood size as a function of threshold length

We know from our historical study of *-ity* that borrowed suffixes in English form emergent self- organizing systems. If words ending in *-ic* or *-ical* formed a simple emergent system, we might predict, based on the overall preponderance 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. We expect to explore other such subsets in further studies. For example, we know that *-ity* and *-ness* are rivals and we have a variety of types of evidence showing that *-ness* is more productive, with some evidence that *-ity* is potentiating in a subset of words (words ending in *–ible* and *-able*). We will use the same Google search techniques to explore the set of words ending in either *-ness* or *-ity*, to see whether there are structural commonalities with the set of words that we have explored here. Other areas for future investigation include *-ize* and *-ify*, the triplet *-dom/-hood/-ship* (German *-tum/-heit/-schaft*), and the broader question of suffix ordering in English.

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.

Selected References

Albright, Adam. 2002. The Identification of Bases in Morphological Paradigms. Ph. D. thesis, UCLA.

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

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.

Plag, Ingo. 1999. Morphological productivity : structural constraints in English derivation. Berlin ; New York: Mouton de Gruyter.

Plag, Ingo and Harald Baayen. 2009. Suffix ordering and morphological processing. Language, 85, 106-149.

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