CHAPTER 5

PRODUCTIVITY, BLOCKING, AND LEXICALIZATION

MARK ARONOFF AND MARK LINDSAY

5.1 Introduction

The topic of morphological productivity as it has been conceived in linguistics for the last half-century is treated in greatest detail in Bauer (2001). If our brief discussion here leads the reader to that book, we will have gone a long way to doing our job. In this chapter, though, we also have a different aim, which is to recast the problem of morphological productivity in a different light. Indeed, we aim to show that the term itself may sometimes be less than helpful. We believe that the most interesting and, more importantly, addressable questions in this domain have always involved not the somewhat elusive notion of productivity, but rather competition. Before getting to that point, however, we must address a more fundamental question, one whose conventional response in linguistics has impeded progress in this particular domain, although it has been enormously helpful in the investigation of other areas of language, whether linguistic systems are entirely discrete in nature.

5.2 Is Language Discrete?

The success of modern linguistics has always been rooted in the realization that languages are systems. But what sort of systems? Linguistics historically has been most successful dealing with discrete patterns and so we tend to assume that languages are wholly discrete systems. The analysis of productivity in word formation presents one of the most serious challenges to date to the blanket assertion that all patterns in language are discrete.

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

The first great achievement of historical linguistics in the 19th century, the comparative establishment of the relationships among the members of the Indo-European and Uralic language families, was based on the discreteness or "regularity" of sound change, what were called phonetic laws and compared at the time to the material laws of science. Famously, the major exceptions to Grimm's law in Germanic languages were later subsumed under Verner's law, providing strong confirmation for the methodological assumption of the regularity, exceptionlessness, or discreteness of sound laws.

In the first half of the 20th century, regularity and discreteness again showed great success, this time in the analysis of phonological patterns, the discovery of the phoneme, and the categorical nature of phonological alternations. Phonemic contrasts are famously categorical and even the distribution of allophones is usually taken to be discrete.

The second half of the last century saw the ascendance of syntax. The immediate constituent analysis of Rulon Wells (1947) led quickly to the phrase structure grammar of Chomsky (1957), which formed the foundation for transformational grammar, all discrete systems. All prominent frameworks for syntactic analysis since then have been discrete.

By the 1960s, especially with Chomsky's (1965) distinction between competence and performance, most linguists could presume comfortably that language was rule governed at its core, so that all components of grammar could be assumed in turn to be discrete systems of regular rules. The messy nondiscrete aspects of language could be relegated to matters of performance or the lexicon, which Bloomfield (1933) had already characterized as a list of irregular items.

Morphology is a challenge for any theory of language that is focused on discreteness and regularity, because so much of morphology is neither. The first challenge for morphologists is to figure out how to integrate regular and irregular phenomena. In inflection, the tried and true method of assuming the dominance of regularity that had succeeded since the days of the Neo-grammarians again proved successful. A variety of researchers from Aronoff (1976) to Pinker (1999) to Brown and Hippisley (2012) worked out the idea that irregular items, listed in the lexicon, could preempt or *block* their regular counterparts, which would emerge as defaults when not preempted by the irregulars. So, the English irregular past tense form *sang* blocks the regular form **singed*, which is the product of the default rule for past tense that adds the suffix *-ed* to English verbs.[1] There is even a hierarchy of outright exceptions like *went* instead of **goed*, rules with narrowly specified domains like the ablaut rules that characterize the relations among *sing, sang*, and *sung*, and the default regular rules.

One way to look at these narrow domains is in terms of the scope of the rules or relations that characterize them. A form like *went* is not describable by any synchronic

[1] In actuality, matters are not so simple. Bauer et al. (2013) show that many English irregular verbs show variation between regular and irregular forms, some well-known case being *dived* vs. *dove, lighted* vs *lit*, and *shined* vs. *shone*. For all these, variation is documented from quite early.

05_9780199641642_c05.indd 68 4/1/2014 8:04:04 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

generalization that goes beyond one verb; *went* must simply be listed as the past tense of *go*.[2] A form like *transcended*, by contrast, is most easily thought of as being rule-derived, in the same way as a sentence like this one must be rule-derived. But the few hundred irregular verbs of English can be thought of as either stored or, more palatably for some, characterized by rules, if we assume that these rules simply have narrower scope than the default rule. The value of the rules for the linguist is that they express the generalizations, admittedly limited, that can be extracted from this set of irregular verbs. The potential wider applicability of these rules is revealed in the errors of children and second-language learners, who may produce forms like *brang*, extending the rule for *sing* to *bring*.

If these irregular but not completely unpredictable phenomena could truly be cast purely in terms of increasingly larger domains, then we could call productivity a discrete phenomenon and preserve the claim that the entire core of language, linguistic competence, is discrete. Unfortunately, the tactic fails. It is not just that the set of *sing/sang* verbs is limited to monosyllables ending in the sequences *-ing* and *-ink*. More importantly, not all such monosyllabic verbs succumb to the rule. Consider, for example, the three homophonous verbs *ring (my bell), ring (the city)*, and *wring (out the clothes)*. Each has its own distinct past tense forms: *rang, ringed*, and *wrung*. We might be able to tag *ringed* as an exception to the smaller-domain rule, so that it then falls under that larger-domain default, but we cannot do that with *wrung*, which must be either lexically listed like *went* or marked as exceptionally showing the vowel that we find in *hang/hung* instead of the vowel that is "normal" in irregular verbs ending in *-ing* and *-ink* like *sing* and *sank*. Furthermore, new verbs of this form are invariably subject to the default rule: *clinked, dinged*, and the website *Blinged out Blondes*, from which all things rhinestone are readily available.

Our inability to cast these phenomena in terms of domains leads next to the notion of discrete degrees of productivity. Default rules like *-ed* are *fully productive*: they apply to any verb that they encounter, except for those that are covered by rules of narrower scope. The rules for the past tense of English strong verbs, by contrast, are less than fully productive: they do not apply to every verb that meets their conditions. We can call them *semi-productive*. But now we need to ask ourselves how many of these discrete degrees there are. As Bauer (2001) so eloquently shows in his chapter on degrees of productivity, this question of how many degrees there are leads to a slippery slope that results inevitably in the abandonment of discreteness as a solution to the problem of productivity in word formation.[3] Bauer's catalog of terms that linguists have used for intermediate degrees of productivity includes, besides *semi-productive, semi-active, active* (though not fully productive), and *marginally productive*. We are led in the end to conclude, as Bauer does, that morphological productivity is scalar rather than discrete and that there is no finite number of degrees of productivity for us to name. As with the points on a compass, we may begin by naming four directions (North, East, South, and West) then

[2] The verb *go* has lacked a morphologically related past tense form since earliest Germanic times, at least. The past tense form *went* is from the verb *wend*, which is uncommon in Modern English but has a regular past tense form *wended*.

[3] The claim that productivity is categorical is asserted as recently as Yang (2005).

05_9780199641642_c05.indd 69 4/1/2014 8:04:05 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

add the four intermediate points Northeast, Southwest, Northwest, and Southeast, but soon we find ourselves needing to talk about South Southwest. Eventually we divide the circle of the compass into 360 degrees, each of these is divided into minutes and then seconds, but in the end we give up and admit that the points on the compass, as with any other circle, are numberless.

5.3 Scalar Productivity versus Blocking in Word Formation

The domain of linguistic inquiry in which the scalar nature of morphological productivity emerges most clearly is that of word formation or lexeme formation (Aronoff 1976, 1994). This was the last core aspect of language to be investigated by modern theoretical linguistics, most likely because of its resistance to discrete methods of study. The treatment of the relative productivity of rival English suffix pairs in Aronoff (1976) provides a valuable history lesson. The unconscious assumption underlying the entire discussion is that the difference between the two rivals is scalar rather than discrete, but the intellectual climate of the time made it impossible for this assumption to be made explicit even to the author of the work, as one of us can attest personally. It was only some years later that the author could even begin to formulate it (Aronoff 1983).

Aronoff (1976: 43) attempted to reduce the contest between rival suffixes to what was termed *blocking*, defined there as "the non-occurrence of one form due to the simple existence of another," a definition since subject to much discussion and some revision (Rainer 1988, Bauer 2001). Blocking, understood in that sense, is discrete: one form exists and the other does not. But later research, some of which we discuss in more detail below, has revealed that this discrete definition fails to capture most of the more subtle interactions that we would surely like to subsume under the term. Most notably, as van Marle (1985) and Rainer (1988) observe, we would like to account for the rivalry within pairs (or larger sets) of affixes, not just between pairs of words, as this definition does. Furthermore, when one word blocks another, the blocked word may still occur, sometimes not with the sense that would be assigned to it if it had no rival, again contrary to this simple definition. In a sense, the word may be deflected instead of blocked.

Here are a few simple examples of how a rival word may be deflected rather than simply blocked. Consider the three English affixes *-ness*, *-ce*, and *-cy*. We can see from the three words *pleasantness, elegance*, and *buoyancy* that they can be rivals, each forming abstract nouns from adjectives.[4] We know that *-ness* is overwhelmingly the overall default suffix for forming abstract nouns from the entire domain

[4] There are others, most notably *-ity*, whose competition with *-ness* is the standard example. But *-ity,* like the two other Latinate suffixes mentioned here, is morphologically conditioned. It does not attach to

05_9780199641642_c05.indd 70 4/1/2014 8:04:06 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

of adjectives but each of the other two suffixes can be more productive than *-ness* in restricted domains. While *-cy* is the least common overall, it is the most favored of the three with the few words ending in *-ate: piracy* (**pirace*), *profligacy* (**profligace*), *delegacy* (**delegace*). By contrast, *-ce* is the most productive with words ending in *-ent* and *-ant: diligence, dependence, resistance*. But in neither of these cases can we say that the rival suffixes are always completely blocked. Sometimes, the *-ncy* rival word is more acceptable than its *-nce* counterpart: *incumbency* (with about 134,000 Google hits) vs. ?*incumbence* (with only 230 or so Google hits). The *OED* lists a fair number of *-nce/-ncy* pairs, and asserts that the former expresses more distinctly the sense of action or process, while the latter expresses the sense of quality, state, or condition, citing the pairs *coherence/coherency, persistence/persistency*, and *compliance/compliancy*. What we actually find is no overall generalization but rather that, when both members of any given pair are entrenched, there is often a difference in meaning and the overall less productive *-ncy* member of the pair conveys a more specialized sense. Compare *excellence* with *excellency*. Both words have a long history in English but *excellency* has come to be used largely in honorific expressions like *your excellency*.[5] *Excellentness* is listed in the *OED* as obsolete, though it shows about 10,000 hits on Google, about a quarter of them from fans of *Bill and Ted's Excellent Adventure*, the classic 1989 cult movie. The pair *compliance* and *compliancy*, cited as an illustration in the *OED*, with *compliance* supposed to signify the action or process and *compliancy* the quality, state, or condition, have both become much more popular in the last century than they were in the 19th because of the importance of the problem in modern bureaucracies. We do not find, however, that the *OED* distinction holds at all in real examples. A cursory examination of actual Google citations reveals that *compliancy* has a more technical flavor and is used for foregrounding and naming: "Let the Compliancy Group solve your compliance puzzle." (<compliancy-group.com>). This is what one would expect from the fact that *-ncy* is overall the less productive of the two suffixes (Aronoff 1983). The same is true for *dependence* and *dependency*. Both have *OED* citations dating to the 16th century and many of their senses have overlapped since then. The latter, however, most frequently signifies "A dependent or subordinate place or territory; *esp.* a country or province subject to the control of another of which it does not form an integral part" (*OED* online). This is a highly specialized concrete sense very far from the general abstract noun sense that characterizes either suffix overall. In general, the idea that a given word bearing one of these three suffixes simply blocks its rivals does not begin to do justice to the complex interaction both among the suffixes and within individual pairs of words.

words ending in *-ant* and *-ent*, and so is not germane in this particular case. The overall power of *-ness* is revealed in its lack of morphological or other conditioning.

[5] In case someone is looking for a pattern, the term *eminence* is used as an honorific, but for cardinals of the Catholic Church only, and the expected *eminency* has had little use since the mid 18th century.

05_9780199641642_c05.indd 71 4/1/2014 8:04:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

It is tempting to see the interactions of rival affixes in terms of synonymy avoidance, of which blocking is a form. Our favorite illustration of synonymy avoidance is an old sociophonetic joke.

Q: What's the difference between a vase [veɪz] and a vase [vɑz]?
A: Oh, about a hundred bucks.

If blocking and synonymy avoidance were driving the interaction of rival suffixes, then we would expect the rival suffixes to each develop a distinct meaning over time. Remarkably, they do not. There have been attempts to show that the much-discussed rival suffixes *-ity* and *-ness* are no longer synonymous (Riddle 1985), but in environments where *-ity* is productive, for example after X_V*-able*, as in *sustainability* or *likeability*, it has precisely the same range of meaning as *-ness* does elsewhere. Only where it is less productive does it show a difference and then it is precisely what one finds with all less productively formed words—specialization, technical usage, and naming. The standard example is *productivity* as opposed to *productiveness*. We talk of *productivity indices* and *personal productivity practices*. There is even a machine tool company named *Productivity Inc.* In none of these instances would *productiveness* do. Finding circumstances under which only *productiveness* is acceptable is difficult, though the following definition of the term *artificially busy* from Urban Dictionary (<http://www.urbandictionary.com>) appears to fit the bill: "A state of activity usually reserved for use in the presence of a manager or boss. The activity mimics productiveness without actually serving a purpose." Here *productiveness* refers only to the state of being productive rather than to some formal measure, which is why *productivity* is at least awkward.

In our own work on rival affixes in English over close to forty years, the only robust example of the members of a set of rival affixes becoming differentiated in meaning is the set *-dom*, *-hood*, and *-ship*. Aronoff and Cho (2001) argue that *-ship* has become specialized to distinguish between stage-level and individual-level attributes. But Lieber (2010a) questions even this case. Based on corpus data she concludes that the three suffixes are frequently interchangeable. This leaves us with no real cases of semantic differentiation in English, the language where this theoretical possibility has been most sought after.

Blocking has proven to be much more successful as a technique in accounting for the interaction of rival realizations in inflection, as opposed to word formation (Brown and Hippisley 2012). Even there, though, problematic nondiscrete rivalries can be found. The English comparative and superlative degree of adjectives, for example, may be expressed either by affixation of *-er* and *-est* or periphrastically with the adverb forms *more* and *most*. Early theoretical accounts of the distribution of the two claimed that the affixal form is found with monosyllables and certain disyllables, with the periphrastic form occurring elsewhere (Aronoff 1976). Less cursory investigation (Graziano-King 1999, Graziano-King and Cairns 2005, Boyd 2007, Gonzalez-Diaz 2008, Mondorf 2009) show that in fact the distribution of the two overlaps, resulting in competition in many individual cases. It may be, then, that even in inflectional systems the distribution of rival forms of realization is not inherently discrete but only becomes so over time.

05_9780199641642_c05.indd 72 4/1/2014 8:04:07 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

The study of productivity in word formation over the last quarter century and more has revealed that it is fruitless to conceptualize productivity of word formation in discrete terms, as all or none. Progress in our understanding has been achieved only by assuming that productivity is scalar, entailing the use of statistical methods. Baayen (2003) and Hay and Baayen (2005) provide persuasive extended arguments for this conclusion.

The realization that morphological productivity is a graded phenomenon opens the doors to new and innovative statistical methods. It also allows us to take advantage of the rapidly expanding electronically analyzable data resources that have become available in this period. In the rest of this chapter, we review some of the statistical methods that have been used for quantifying and measuring productivity, along with results. We begin with the best-known electronic corpus-based method, Harald Baayen's measures of productivity based on *hapax legomena*, words that only occur once in a corpus.[6] We then move to two methods that we have used ourselves. The first takes advantage of digital versions of the *Oxford English Dictionary*, and allows one to trace the productivity of affixes over time. This method was first developed by Anshen and Aronoff (1999). The last method takes advantage of the vast and ever-expanding virtual corpora made possible by the World Wide Web.

5.4 *Hapax Legomena*

Linguists have struggled to precisely define what productivity is; quantifying and measuring productivity is, therefore, also problematic. The most useful measures of productivity over the past twenty years have come from the work of Baayen and colleagues, particularly Baayen (1992) and Baayen (1993). These measures, *P* and *P**, center on the notion of the *hapax legomenon*, or a word that occurs only once in a corpus. Baayen's underlying assumption is that there is a strong relationship between hapaxes (as they have come to be called instead of the "proper" Greek plural *hapax legomena*) and productivity.

Baayen's first measure is *P*, which Baayen (1993) calls the Category-Conditioned Degree of Productivity. For a given affix, P is defined as:

$$P = n_1/N$$

where n_1 represents the total number of hapaxes containing the affix, and *N* represents the total number of tokens containing the affix. This measures the "growth rate" (Baayen 1992) of the affix: the probability that an encounter with a word containing the affix reveals a new type.

[6] The term *hapax legomenon* 'read once', sometimes plain *hapax*, is Greek and originates in the scholarly study of the Bible, where the meaning of a word that only occurred once in the received text might be especially difficult to discern, making such words of special interest.

05_9780199641642_c05.indd 73 4/1/2014 8:04:08 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

Ideally, n_1 would precisely represent all individual word types in a corpus that were productively derived, regardless of the number of times the word occurs in the corpus; however, it is certainly not feasible, and probably impossible, to systematically test whether a given token in a corpus was created productively or came from that speaker's lexicon. Both *P* and *P** (which we will discuss shortly) must rely on the assumption that hapaxes are a good representative of productive word formation; indeed, Baayen (1993:189) explains that the probability of encountering neologisms "is measured indirectly" via the counting of hapaxes, and that not all hapaxes are neologisms, and vice versa. Given this, it is crucial that it be true that, if a token occurs only once in a corpus, it is proportionately more likely to be productively formed, and, conversely, if a word is productively formed, it is proportionately more likely to occur only once in a corpus. Intuitively, this seems like a sensible assumption, but it is difficult to prove; to do so would require, at the very least, a precise, agreed-upon set of criteria for categorically judging an occurrence of a word to be productively formed or not productively formed. Paradoxically, the need for measurements like *P* and *P** arise precisely because this cannot be accomplished. Instead, Baayen supports the use of *P* and *P** as measures of productivity by analyzing their predictions: do the measurements produced by these methods yield results that correlate with our intuitions about the affixes in question?

Baayen (1992) assesses the validity of *P* by comparing rival suffixes such as English *-ity* and *-ness* using the CELEX database. Of the two, *-ness* is qualitatively regarded as much more productive than *-ity*, although there are a large number of established *-ity* types.

As shown in Table 5.1, Baayen's *P* measure produces a value of 0.0007 for *-ity* and 0.0044 for *-ness*, even as the number of types (405 and 497, respectively) is very close between the two. Further, both *P* values are higher than the *P* value of 0.0001 for simplex nouns in the corpus (which, by definition, are not formed through productive processes).

To evaluate global productivity, Baayen suggests considering both *P* and *V* together, where *V* is the number of individual word types in the corpus (the vocabulary size). Differences in *V* reflect the extent to which relevant base words have been used, while differences in *P* relate to differences in extent that remaining base words can be used to create neologisms.

Table 5.1 Comparing the productivity of *-ity* and *-ness*

affix	Tokens	types	Hapaxes	P
simplex nouns	2142828	5543	128	0.0001
-ity	42252	405	29	0.0007
-ness	17481	497	77	0.0044

Source: Adapted from Table 2 in Baayen 1992.

05_9780199641642_c05.indd 74 4/1/2014 8:04:09 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

Table 5.2 ***P**** **value comparison**

category	$P^* \cdot$ h1 (a.k.a. hapaxes)	qualitative judgment of productivity
simplex nouns	256	–
-ness	77	+
-ation	47	+
-er	40	+
-ity	29	+
-ment	9	±
-ian	4	±
-ism	4	+
-al	3	±
-ee	2	±

Source: Adapted from Table 3 in Baayen 1993.

Baayen's second measure, the Hapax-conditioned Degree of Productivity, *P**, is defined as the following:

$$P^* = n_1/h_1$$

Again, n_1 is the total number of hapaxes with a given affix, while h_1 is the total number of hapaxes across all types in the corpus. This measure predicts the likelihood that any new word that one encounters will contain the affix, which, according to Baayen (1993: 193) "can also be viewed as measuring the relative contribution of a given morphological category to the overall vocabulary growth."

*P** is tested in Baayen (1993) in similar fashion to *P*, by judging its predictions. Because h_1 is the count of total hapaxes in the corpus, the denominator in the expression n_1/h_1 is the same for all suffixes; therefore, when comparing *P** values in a given corpus, we are, in effect, simply comparing the number of hapaxes occurring for each suffix.

Returning to *-ity* and *-ness*, we see in Table 5.2 that there are 29 *-ity* hapaxes and 77 *-ness* hapaxes, which is again in line with our intuition that *-ness* is more productive. With *P**, however, it is not possible to compare values to a baseline of simplex nouns, as the *P** value for this category is very high.

Both *P* and *P** measurements are dependent on the size (*N*) of the corpus. The number of hapaxes in a corpus is a decreasing function of *N*; ultimately, the rate of increase in the number of hapaxes slows as the size of the corpus increases. This means that comparing measurements across corpora is problematic.

05_9780199641642_c05.indd 75 4/1/2014 8:04:10 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

Hay and Baayen (2002) show a link between parsing and productivity—namely, Baayen's *P* measurement of productivity. For words containing a given affix, Hay and Baayen plot the frequency of the derived forms against the frequency of bases of those forms. In forms where $x = y$, the frequency of the base is equal to the frequency of the derived form; Hay and Baayen call the line $x = y$ the *parsing line*. Those forms that are plotted below the parsing line are words that are more frequent than their bases (e.g. *illegible* is more frequent than *legible*). Forms above the parsing line have bases that are more frequent than the derived forms. Hay and Baayen claim that those words falling below the parsing line are more likely to be accessed as whole words, rather than component parts, while words above the line are more likely to be decomposed and the affixes, therefore, used productively. They then calculate the *parsing ratio* for a given affix, that is, the proportion of words that appear above the parsing line. Hay and Baayen find that *P* is a strong predictor of this parsing ratio, providing further support for the validity of *P* as a measurement of productivity.

5.5 Dictionaries

Detailed historical dictionaries, such as the *Oxford English Dictionary* (*OED*), can provide a rough, but nonetheless insightful, diachronic survey of productivity. They allow us to address a different question from most studies of morphological productivity. Rather than what it means for a given morphological structure to be more or less productive than another, they allow us to study how a given morphological structure has become more or less productive over time.

Any dictionary is subject to the biases of the editors. Dictionaries of standard written languages tend to favor works of well-regarded authors as major sources of citations. Thus, one can neither assume that all "existing" words are in the dictionary, nor that all words in the dictionary are currently "existing" words. However, it is probably impossible to compile such a list.

> Only in the kind of ideal world that contains ideal speaker-listeners can we hope to find a list of existing words. It follows that the methodologically practical assumption of the equivalence of the word-list of any reference work or set of reference works and the set of existing words is inevitably flawed. (Bauer 2001: 36)

Anshen and Aronoff (1999) use the *OED* on CD-ROM to investigate the birth and death of borrowed suffixes in English. Using the software's advanced searching tools, one can search for words matching certain criteria, such as all words ending in the suffix *-ity*. Each word's entry contains (among other things) definitions, etymological information, and citations. The date of first citation can be used as an approximate indicator of when a word came into use, while the etymology makes it possible to determine the likely

05_9780199641642_c05.indd 76 4/1/2014 8:04:10 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

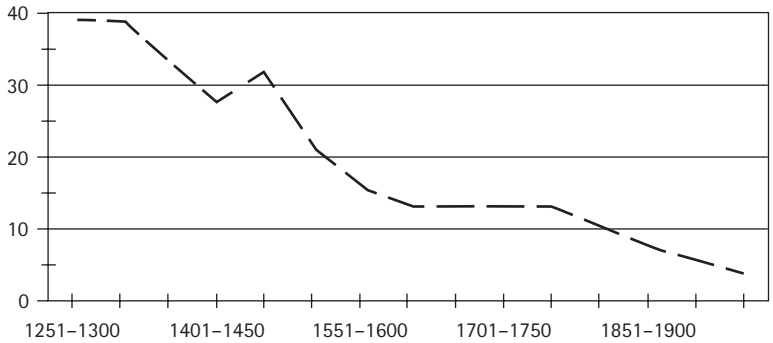

FIGURE 5.1 French borrowings as a percentage of all new words

Source: from Anshen and Aronoff 1999.

language of origin. The latter piece of information can be used to categorize a word as being borrowed or native to English.[7]

Grouping these dates of first citation into bins by century or half-century, one can graph the number of words cited for the first time in each time period, giving an approximation of how productive an affix has been over time. In Figure 5.1 and Figure 5.2, Anshen and Aronoff show the birth of *-ity* as a productive suffixation pattern in English. Figure 5.1 illustrates the gradual decline in the borrowing of French words into English. Over the same time period, we see in Figure 5.2 that an increasing percentage of new *-ity* words were being derived in English. Indeed, during the 19th century, 937 new *-ity* words were derived, while only 35 were borrowed.

We see a gradual increase in *-ity* derivations, even as borrowings decrease; this is the birth of *-ity* as a productive suffixation pattern in English. Once a sufficient number of borrowings had entered into the language, speakers could then abstract out the suffix and generalize a pattern.

Anshen and Aronoff then use this same method to compare the differing fates of productive *-ment* and *-ity*. While *-ity* is a productive suffix today, *-ment* has all but fallen out of use, except as a fossilized component of established words.

In Table 5.3, Anshen and Aronoff track the number of new derived *-ity* and *-ment* forms entering English.

Here we see a strong decline in new *-ment* words beginning in the 17th century, while *-ity* generally holds strong into the present day. This decline in *-ment* derivations coincides with a change in the number of new verbs entering into English, as shown in Figure 5.3. Since *-ment* is dependent on new verbs for productivity, a decline in potential

[7] Defining the language of origin for a word is by no means trivial, and the origins chosen by the *OED* are subject to interpretation. For example, if an affixed word's first citation in English comes later than the first citation of the same word in French, this does not guarantee that the English speaker did not simply derive the same word natively. However, it may be impossible to determine such a thing definitively.

05_9780199641642_c05.indd 77 4/1/2014 8:04:11 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

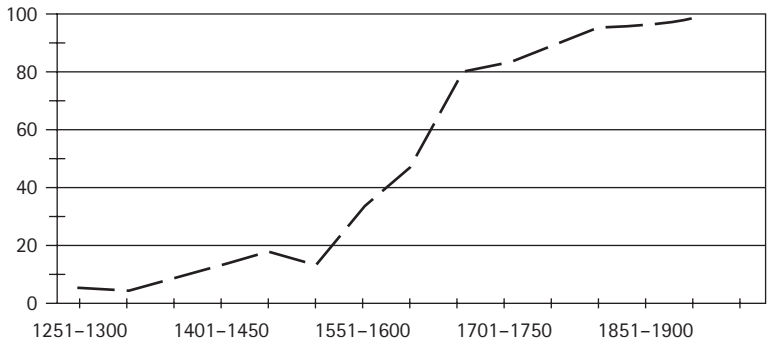

FIGURE 5.2 Derived *-ity* as a percentage of all *-ity* words

Source: from Anshen and Aronoff 1999.

Table 5.3 Derived forms for *–ment* and *–ity*

Half-centuries	Derived *-ment*	Derived *-ity*
1251–1300	6	1
1301–1350	10	1
1351–1400	19	11
1401–1450	15	11
1451–1500	37	16
1501–1550	60	22
1551–1600	174	64
1601–1650	217	206
1651–1700	76	241
1701–1750	40	108
1751–1800	37	177
1801–1850	158	435
1851–1900	142	502
1901–1950	26	298
1951–2000	4	179

Source: from Anshen and Aronoff 1999.

hosts should have an impact on its performance as a productive pattern; on the other hand, *-ity* relies on adjectives, which enter English in greater numbers during the same time period, and it thrives productively.

Lindsay and Aronoff (2013) improve on this claim by normalizing the *OED* data to account for the variable amount of source material from century to century. After

05_9780199641642_c05.indd 78 4/1/2014 8:04:13 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

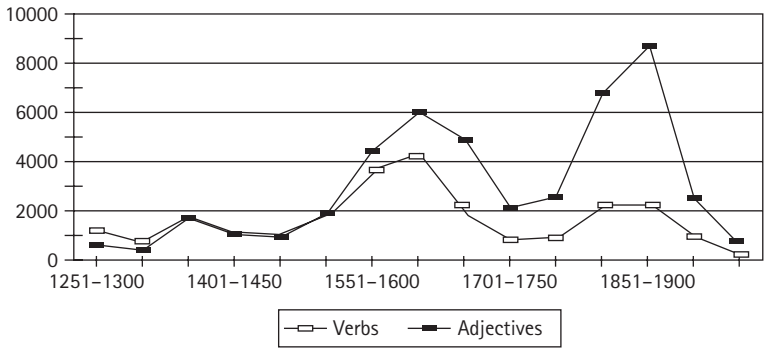

FIGURE 5.3 Number of new English verbs and adjectives

Source: from Aronoff and Anshen 1999.

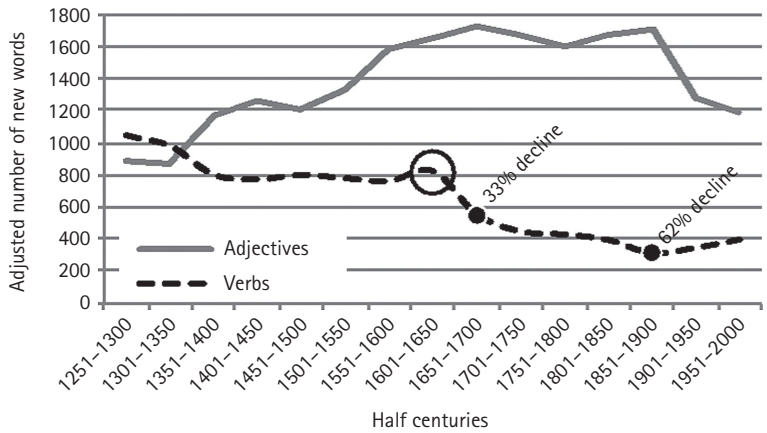

FIGURE 5.4 New adjectives and verbs entering English, showing a rapid decline in the relative number of new verbs beginning in the 1600s.

Figure 5.3 is adjusted,[8] the difference between new verbs and new adjectives becomes much clearer (Figure 5.4).

Likewise, the divergence between *-ity*'s productivity and *-ment*'s productivity is also much more pronounced (Figure 5.5, compared to Table 5.3):

Thus, we see that historical dictionaries like the *OED* provide a practical means of analyzing the relative productivity of suffixation patterns diachronically. This information can be used to track the emergence (and death) of productive processes and identify factors that may have influenced their fates.

[8] 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:

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

05_9780199641642_c05.indd 79 4/1/2014 8:04:15 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

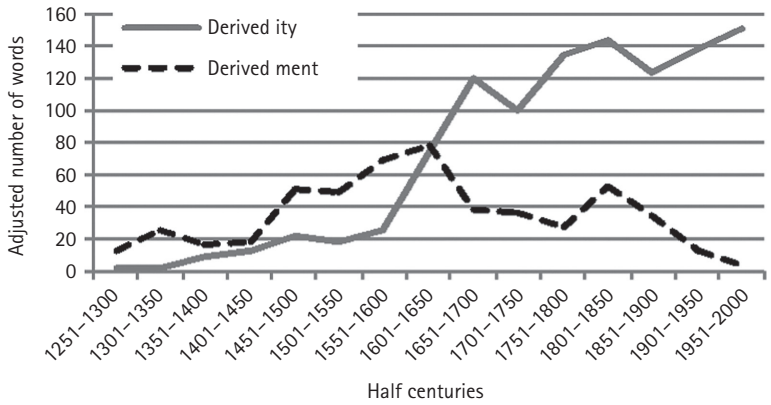

FIGURE 5.5 New derivations of *-ity* versus *-ment* over the past 750 years.

5.6 The World Wide Web and Productive Rivalries

Rivals *-ic* and *-ical* are both productive today in spite of their mutual dependence on the same pool of stems and no distinguishing semantic differences.[9] Why do some rival patterns seem to stabilize and coexist, while others do not? Lindsay and Aronoff (2013) view languages as self-organizing in a manner similar to biological systems; languages are complex, continuous systems that change through numerous smaller interactions, a phenomenon known as glossogenetic evolution (Hurford 1990, also discussed in Steels 1997 and Fitch 2010, among others).

Lindsay and Aronoff (2013) also use another resource, the world wide web, to examine synchronic productivity of suffix rivalries in English. To accomplish this, they use statistical estimates from search engine results—in this case, the Google Search engine, using the Google Search API.[10]

One must be cautious when incorporating Google Search's Estimated Total Matches (ETM) 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

[9] While word pairs like *electric* and *electrical* have different meanings, these differences are not generalizable; the difference between these words bears no resemblance to the difference between e.g. *historic* and *historical*.

[10] As of January 2012, the Google Search API has been discontinued for all purposes, including academic research. Querying Google for results is still technically possible, but much less practical, due to the constraints made by Google on query frequency.

05_9780199641642_c05.indd 80 4/1/2014 8:04:16 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

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. A broad investigation of suffixes mitigates many of these concerns when dealing with single words, regular inflection patterns, and a large number of stems (Lindsay and Aronoff 2013).

To gather the data, first, a list of suitable words must be generated in order to feed them into Google Search.[11] Using basic regular expression matching (along with some manual filtering), we can identify all words ending in either *-ic* or *-ical* in Webster's 2nd International dictionary.[12] The suffixes of these words are then stripped off, leaving bare stems; duplicate stems are discarded. In the case of *-ic*/*-ical*, this yielded 11,966 unique stems that take *-ic*, *-ical*, or both suffixes.

Next, each stem-suffix combination is automatically queried in Google Search as a literal string (e.g. *biolog* + *ic, biolog* + *ical*) and the ETM value is returned and recorded in a database.

ETM values are then compared and analyzed. In Tables 5.4 and 5.5, we see a sample of ETM values for various *-ic*/*-ical* pairs:

In some cases, both *-ic* and *-ical* have a substantial number of tokens (Table 5.4), though in the majority of cases, one suffix yielded far more results than the other (Table 5.5). Overall, 88.5% of pairs differed by at least one order of magnitude.

By comparing ETM values for each form for a given stem (e.g. *biolog-ic* and *biolog-ical*), the assumption is that the more productive suffix will tend, over a large number of comparisons, to have a higher ETM value more often than the less productive suffix.

Between *-ic* and *-ical*, *-ic* was found to be the "winner" in 10,613 out of 11,966 pairs. However, *-ic* was not preferred in all domains. Lindsay and Aronoff systematically examined all neighborhoods appearing on the right-edge of the list of stems. For example, one could look at the final letter of each stem (neighborhood length 1) and find that there are 4166 stems ending in *t*, or look at the final two letters (neighborhood length 2) to see that there are 1129 stems ending in *st*. If one continues this process for all possible combinations, generalizations begin to emerge.

Naturally, there is clustering in certain neighborhoods, with the largest groups usually (but not always) coinciding with traditional morpheme boundaries (e.g. *graph*). Most of these groups also favor *-ic* over *-ical*; however, there is one significant exception: stems ending in *olog* (of which there are 475) favor *-ical* over *-ic* by a ratio of 6.42 to 1 (Table 5.6).

[11] At one time, it was possible to use regular expression matching in search engine queries; for example, Hathout and Tanguy (2002) created the WebAffix tool for these types of advanced queries using the AltaVista search engine. Unfortunately, advanced regex matching, if present at all, is severely restricted in present-day search engines. Thus, querying must be done without the direct use of regex matching.

[12] This was chosen as a source, in part, because it was freely available in digital form.

05_9780199641642_c05.indd 81 4/1/2014 8:04:17 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

Table 5.4 Sample Google ETM counts for high-frequency doublets

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

Source: from Lindsay and Aronoff 2013.

Table 5.5 Sample Google ETM counts for high-frequency singletons

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}

Source: from Lindsay and Aronoff 2013.

Table 5.6 *-ical* is productive in stems ending in *olog*

	Total stems	Ratio	*-olog* stems	Ratio
Favoring -ic	*10613*	*7.84*	74	1
Favoring -ical	1353	1	*401*	*6.42*
Total	11966		475	

Source: from Lindsay and Aronoff 2013.

Here we see a possible explanation for the ability of rivals *-ic* and *-ical* to coexist productively in a competitive ecosystem. While *-ic* is strongly preferred overall, *-ical* is not being driven out of the system because there exists a coherent subdomain of sufficient

05_9780199641642_c05.indd 82 4/1/2014 8:04:18 PM

OUP UNCORRECTED PROOF – FIRSTPROOFS, Tue Apr 01 2014, NEWGEN

size in which *-ical* is preferred. Without this niche, we should not expect *-ical* to survive as a productive entity (see Lindsay and Aronoff 2013 for further discussion).

While Google Search estimates cannot be employed for all linguistic (or even all morphological) investigations, in certain applications these data can provide valuable insight. Although the research program for Google Search has been terminated, the Google Books corpus remains freely available. Version 2 of the English language corpus (released in 2012) provides information on the date of citation, as well as part of speech tagging, on millions of books and publications.

05_9780199641642_c05.indd 83 4/1/2014 8:04:18 PM