Self-organization in the spelling of English suffixes: The emergence of culture out of anarchy[*]

Kristian Berg, University of Oldenburg
Mark Aronoff, Stony Brook University

COMMENTS WELCOME

[*]This paper was supported by the German Science Foundation (DFG) project "Prinzipien der Wortschreibung im Deutschen und Englischen" ('Principles of word spelling in German and English') and a DAAD post-doctoral scholarship to Kristian Berg held at Stony Brook.

1 Introduction

English culture is famously unprincipled. For starters, England has never had a constitution. Magna Carta, signed eight hundred years ago, is sometimes held up as an English bill of rights, but the only ones whose rights it protected were barons. Since then and before then, England and the rest of Great Britain have gotten along very well without a constitution of any sort, thank you, making the country a very conspicuous exception among the constitutionally based nations that dominate the modern world. Even when it comes to simple written law, the English are sorely lacking. Not just England, but almost all of its former colonies stand out in following common law, which is governed largely by precedent rather than by written statute. No constitution, no laws, what kind of culture is that?

No spelling or grammar rules either. The English language has just as staunchly and just as successfully resisted statutory regulation as has its government and its legal system. The founding of the Accademia della Crusca in Florence in 1582 and the Académie Française in 1735, both devoted to overseeing language, led such British scientific and literary luminaries as Robert Boyle, Jonathan Swift, and Joseph Addison to propose a similar governing body for the English language. Strongly opposed by Samuel Johnson on the grounds of "English liberty," (Martin 2008, p. 197), the idea quickly fell out of fashion, leaving the language with no government or police.[1] Nonetheless, as we will show, just like a spoken language, English spelling has arrived at a system despite the lack of any overt guidance.[2]

It has been suggested that the English language has overwhelmed the globe because it has no one to police it. No nation or authority of any sort owns English and neither the United Kingdom nor the United States has an official national language. This permits anyone in the world to use English as they will, making up their own words and constructions with no official interference. There is not even an authority anywhere in the world governing how English is spelled. Present-day English spelling varies from country to country, enforced only by local editorial practice, which may differ from one publisher or organization to another.

English spelling, in any of its current incarnations, appears to be as lawless as it is ungoverned, anarchy run amok. It is notoriously unphonetic, rivaled in that regard only by French spelling, and examples abound of the same sound spelled in different

[1] A British Academy was eventually chartered by royal decree in 1902. According to its own official history, the academy "was first proposed in 1899 in order that Britain could be represented at a meeting of European and American academies" (because it had none!). This academy, however, has never had any jurisdiction over the English language.

[2] The analogy between English spelling and English common law is not entirely apt. Common law rests on court precedent (*stare decisis*), which depends on having access to the recorded decisions of individual judges within a legal hierarchy. For spelling we have neither courts nor records to guide us; one aspect of the analogy that does hold is that we can only understand the current system through its history.

ways. The major consistency is lexical, in the spelling of individual words: a given word will be spelled in one way within a given tradition. Word spellings may differ from one tradition to another, sometimes in complex ways: US spelling uses *judgment* while the usual British spelling is *judgement*, except that the British use *judgment* to specifically denote a judicial decision. Homophones that are not also homographs, distinct words that sound the same but are spelled differently, provide a striking example of lexical consistency. Each of the three identical-sounding words *pare, pair,* and *pear* has a consistent but distinct spelling. Among alphabetic writing systems, the use of distinct spellings to differentiate words visually is most pronounced in English and French. The origins of this strategy are unclear, though both systems were fixed at about the same time, in the half-century or so after 1650. The downside of the strategy is that it wreaks havoc with sound-spelling correspondences, as most critics of French and English spelling have noted.

This article is a study of the middle ground between the spelling of sounds and the spelling of words: the spelling of affixes, specifically suffixes. We will show that the spelling of any given English suffix is quite consistent, despite the absence of any external authority making it so. This observation is not entirely new. It has often been remarked (Carney 1994: 18ff.) that the two most common English inflectional affixes, <-*s*> and <-*ed*>, while they vary in form depending on their phonological environments, do not vary in spelling. <-*s*> can be pronounced as either [s], [əz], or the default [z], depending on the sound at the end of the word to which it is attached, while <-*ed*> is either [t], [əd], or [d]. In both cases the spelling remains the same despite the different pronunciations (which are admittedly predictable). The suffix <-*s*> is polyfunctional. It can represent either the plural of nouns (*cats*), the third person singular present of verbs (*tends*), or the possessive of nouns (*men's*), but the possessive is orthographically distinct, carrying an apostrophe before it, so that the plural *cats* and the possessive *cat's* differ from one another in form, though the plural noun *dogs* and the singular verb *dogs* are indistinguishable.[3]

Chomsky and Halle (1968) have famously noted that lexical consistency extends even to derived words, citing such sets as {*sane, sanity*}, {*sign, signify*}, and {*electric, electricity, electrician*}. In each case, the spelling of the base word remains the same throughout, despite the phonological changes consequent on suffixation. Although they both signal some sort of constant, there is a subtle difference between the constant spelling of lexemic stems despite differences in pronunciation in cases like *electric/electricity/electrician* and the constant spelling of suffixes that is the object of our study here. The lexemes are spelled in the same way despite differences in pronunciation in different environments. A given suffix is spelled the same way across different words that contain it and this constant spelling differs from that of the same phoneme sequence in instance where this phoneme sequence does not represent the suffix.

[3] The regular genitive plural marker is <-'>, a silent apostrophe, since the genitive suffix does not occur after the plural [-*s*] on account of haplology. This may be the only case of a true zero marker in English spelling.

Put another way, the spelling system follows the general pattern of distinguishing homophones, peculiar to English and French spelling, which we saw already in examples like *pair/pear/pare*, but it extends the pattern beyond words to word endings. As mentioned above and as we will show, affixes are spelled differently from homophonous sequences that that happen to fall at the ends of lexical words. The system spells the denominal adjectival suffix <ous> consistently, while all other words that end in the same sequence are spelled differently. The words *nervous, office,* and *tennis* all end in the phonological sequence [əs] but only *nervous* contains the suffix.

We have arrived at the two larger questions that we address here. First, to what extent does present-day English spelling call attention to the spelling of individual affixes beyond the two inflectional suffixes? Second, how did it reach its current state? The answer to the second question bears on the larger and much more interesting general question of how a system can emerge in the absence of any stated principles or guiding hand, how English spelling, like English common law, came to take on the shape that it has today. We will show that affixal constancy is characteristic of the current state of English spelling. But matters were not always so. Overall, our study reveals that Matthew Arnold (1869) was wrong in contrasting culture and anarchy. Culture, at least in this case, arose out of anarchy.

From a much wider perspective, the emergence of systematicity in English spelling is another example of the workings of competition, the struggle for existence. English spelling from the Middle English period through the end of the seventeenth century was unsettled because history had provided numerous ways to spell the same word. Queen Elizabeth, for example, who left a large legacy of autograph correspondence, had a single spelling for only half the words that she used in the documents that have been preserved (707 out of 1389), though she was not entirely unsystematic: 523 of the remaining 682 words showed only two variants (Evans 2012). The history of English spelling since 1600 shows what happened to many of the available variants: their distribution became lexically fixed, for both lexemes and suffixes.

In ecological terms, each variant spelling has found its niche. What defines these niches, though, is not spelling. Spelling simply fills niches that are made possible by the morphology of the language. As one of us once noted, "written language is a product of linguistic awareness, the objectification of spoken language. Any orthography must therefore involve a linguistic theory," albeit an implicit one (Aronoff 1985, p. 28). Our work shows a clear implicit linguistic theory lying behind English orthography: a language contains lexicalized units, both free-standing lexemes and affixes. English spelling emphasizes these lexicalized units at the cost of the consistent representation of phonemic units that the original alphabet highlighted, which persists in most alphabetic systems to this day (Saussure 1916).[4]

[4] Unlike Chomsky and Halle (1968) we remain silent on whether English spelling is or is not a good writing system, compared to a more phonologically consistent

This article has two major aims: to show that present-day English writing refers to morphology in subtle ways and to show how the system we witness today came to take its current form. We first investigate whether the relation between the written form of a suffix and the occurrence of that suffix is reliable. Thus, if a word ends in <ous>, what are the chances that <ous> is the suffix that derives adjectives from nouns? We then determine whether this relation is an accidental reflex of phonology or whether it is exclusively graphemic/morphological? How many words end in the phonological sequence [əs] and could potentially be spelled with final <ous>, but are not? Compare, for example, <nervous> and the noun <service>, which does not contain the suffix in question. Why isn't the latter *<servous>? Do any of these words form classes of their own? Is there an overall pattern?

Our overall linguistic approach is word based (Matthews 1972, Aronoff 1976, Anderson 1992, Blevins 2013). This means that we pay attention to whole words and to at least somewhat productive word-based morphology. We treat the word <service>, for example, as a morphological whole and not as consisting of a root or stem and a suffix, because there is no productive suffix [əs] in English that forms nouns from verbs. By contrast, <nervous> is analyzable because the suffix [əs]/<ous> does form adjectives from nouns (Marchand 1969). Because the morphology is word-based, in our framework the word <nervous> both stands as a lexeme in its own right and contains the suffix we are interested in. We can have our cake and eat it too.

As for graphemics, we treat it as a system in its own right; it has regular correspondences to other levels of linguistic description, but it needs to be analyzed autonomously, without recourse to phonology or morphology, before we can begin to ask if, and then how, graphemic units correspond to phonological or morphological units. In structural linguistics, there is a tradition dating to at least Saussure ([1915] 1959) to regard writing as secondary to spoken language. We take no stance on that question. We are interested solely in the extent to which correspondences can be found between regularities in the writing system and those in the spoken language.

2 Methodology

---

system. It is entirely possible that a writing system like that of English that calls attention to lexical units is more usable than one that is more phonologically grounded. It is clearly less easily learned (Treiman & Kessler 2014). It also makes sense to distinguish different ways in which an alphabet can be usable: Phonologically consistent spellings are probably more useful if you're reading out loud, but spellings that prioritize standardized lexical access may be more useful for silent reading.

2.1 Synchronic investigation

The basis for the synchronic investigation is the CELEX database (Baayen et al. 1995). This database contains 52,447 English lemmas. Each lemma comes with a graphemic and phonological form, and many also contain information about the morphological structure and lexical category. We will use this corpus to answer the questions about the relation between spelling and morphology.

The way CELEX is structured creates four problems for our investigation. First, for 8,490 entries (16% of the total), the morphological structure is dubbed "obscure" (words like *amorous*), "irrelevant" (e.g. *arctic*), "may include a 'root'", (e.g. *brandy*), or "undetermined" (e.g. *causerie*). As a consequence, these words do not have a lexical category assigned to them. To solve this problem, we use the Oxford English Dictionary (OED) to look up missing word categories and then add them to the database.

Second, conversions between word categories are separate lexical entries in CELEX: the verb *run* is one entry with its own graphemic, phonological and morpho-syntactic information, and the noun *run* is another one. For our goals, this is problematic. We want to be able to determine what ratio of words with a given graphemic ending is a possible member of a lexical category; e.g. how many words that end in <ous> are potential adjectives? To do so, we need to treat two (or more) entries with the same form as just that – one form, and note which possible word categories this form can appear in.

Third, there are many cases where a given lemma also occurs as a part of another, more complex lemma. If, for example, we are interested in words ending in [ɨk], we find the lemma *music*, but we also find *canned music, country music, chamber music, incidental music, piped music, programme music, sheet music,* and *soul music*. We would like to treat all these cases as instances of one lemma, however, namely *music*, which is modified by a second stem. Accordingly, we exclude all entries that consist of more than one stem.

Fourth and maybe most importantly, some of the phonological transcriptions appear to be inconsistent when it comes to suffixes and word endings, or more generally: to reduced syllables. Take the phonological minimal pair *nervous*/*service* for example. While *nervous* is transcribed in CELEX with a schwa in the second syllable (['nɜ.vəs]), *service* has a near-front near-close unrounded vowel (['sɜ.vɪs]). This distinction is not justified phonetically. As Flemming and Johnson (2007) show, there does not seem to be a difference in the realization of non-final reduced vowels (although there is a difference if the vowel is stem final as in the famous pair *Rosa's/roses*). Flemming and Johnson propose to transcribe all reduced vowels in non-final position as close central unrounded vowels ([ɨ]), and we follow this suggestion. The phonological transcriptions in CELEX are modified accordingly: *nervous* and *service* are now ['nɜ.vɨs] and ['sɜ.vɨs], respectively.

Additionally, we exclude some entries from the analysis. The first group are proper names. The spelling of proper names can be much more idiosyncratic than the rest of the lexicon (cf. e.g. Carney 1994: 443ff.). The second group are abbreviations like *anon*. or *usu*. (for *anonymous* and *usually*, respectively). As abbreviations, they do not have a corresponding phonological structure (other than that of the full word they refer to).

We have investigated four derivational suffixes.[5] Each of these suffixes consists of an unstressed syllable rhyme that can be spelled in several ways in English. Our question is whether one of the available spellings for each of these phonological rhymes is particularly associated with the suffix rather than with a morphologically unanalyzable word-final unstressed rhyme. For each suffix *(ous, -ic, -al, -y*)[6], we did the following:

1. We determined how many words in CELEX end with the graphemic form of the suffix and could be confused with it. Take <-*al*> for example: How many words are there that end with these letters? Lemmas like *pal* have to be excluded, since <-*al*> in this word does not run the risk of being interpreted as a suffix. The following well-formedness constraint applies to what we call the "stem" (the lemma stripped of its word ending): It contains at least one graphemically closed syllable (e.g. <leth> in <lethal>, but not <re> in <real> or <vi> in <vial>). Although this constraints may seem unmotivated and ad-hoc, there is actually evidence for it from reading psychology (cf. e.g. Taft 1979).

2. Of these, we determined the ratio of words that can be argued to bear the suffix. Take -*ic*, for example: This suffix forms adjectives. Any word that ends in <ic> and is a potential adjective in this sense bears the suffix. Words with final <ic> that are not adjectives do not contain the suffix (e.g. *panic*). In more theoretical terms, we use the output of word-based word-formation rules (Aronoff 1976, 1994) as constraints that determine group membership.

3. We determined how many words could **potentially** be spelled like those ending in the suffix, judging solely from their phonology. For this, we took the phonological form of the suffix as a basis and searched for all words that end with this phonological sequence. Note that by phonology we do not only mean the

[5] Berg et al. (2014) provides a similar analysis of the two most common inflectional suffixes.

[6] We have chosen to use (italic) spelling without angled brackets for the suffixes in question, in order to distinguish the suffixes from the letter strings and the phonological sequences that they each correspond to. This use of spelling rather than phonological notation is traditional in the word formation literature (e.g. Jespersen and Marchand). A given letter string may or may not instantiate the corresponding suffix. So, -*ic* designates the adjectival suffix but <-ic> designates the word final letter string and [ɨk] the phonological sequence. A word like *music* contains the letter string and the phonological sequence but not the suffix.

occurrence of some word-final phonemes, but also prosodic patterns: The suffix [ɨk] is never stressed, and in each word where it occurs, there is always at least one more syllable. Neither *sick* nor *sic* would qualify by this definition.

4. We determined the pattern of distribution and correlated different spellings with different morphological features. For example, the words in [ɨk] fall into two classes: those that can be adjectives (i.e. those that bear the suffix), and those that can not. As it happens, this particular functional distinction is mirrored closely by spelling: Almost all words that can be adjectives end in <ic>, while almost no word that cannot be an adjective does.

2.2 Diachronic investigation

To determine how the spelling system we find today evolved we use the Helsinki corpus (The Helsinki Corpus of English Texts 1991). This corpus is a collection of extracts from continuous texts that date from between 750 and 1700 and contains 1,572,800 words. It is divided into eleven time spans of 70 to 100 years, which subdivide the traditional historical periods of English:

historical period	**time span**	**word count**
Old English	-850	2,190
	850-950	92,050
	950-1050	251,630
	1050-1150	67,380
Middle English	1150-1250	113,010
	1250-1350	97,480
	1350-1420	184,230
	1420-1500	213,850
Early Modern English	1500-1570	190,160
	1570-1640	189,800
	1640-1710	171,040

Table 1: Overview of the relation between traditional periods of English and time spans in the Helsinki corpus, plus the number of words in each time span.

The vast majority of the texts included in the corpus are public in nature, such as handbooks, treatises, biographies and proceedings, but there are also some private letters and diaries (for this and the following, cf. The Helsinki Corpus of English Texts 1991). The corpus is not morphologically or syntactically annotated, and it does not contain a tier with normalized orthography. The fact that e.g. <cite>,

<cittee>, <city> (among others) are all spellings of one lexeme (*city*), is information that is not contained in the corpus but which must be gathered manually.

One potential problem with the Helsinki corpus is that it was not explicitly compiled for the investigation of spelling. Orthographic faithfulness was, in other words, not a priority. While the Old English texts had previously been digitized (for the *Dictionary of Old English* project at the University of Toronto), the other texts were keyed in from editions or early imprints. When possible, modernized editions were avoided, and if several editions existed for one text, they were compared to find the most reliable one (Merja Kytö, p.c.).

Diachronically, it is of interest how the system of suffixes we find today came into existence. That means we want to find every instance of every word spelled with a given suffix today, and investigate the formal changes to that suffix over the time span covered by the Helsinki corpus. For a list of possible spellings, we used the OED online. For example, the OED gives (among others) the following spelling variants for *-ous*: <ose>, <ows>, <is>, <owse>, <ys>, <es>, <ouse>, <us>, <ous>. All words that end in one of these forms were therefore searched in the Helsinki corpus. As a first approximation, we used a constraint on word length: The word stripped of the word ending had to be at least three letters long. This step excludes hits like *this* and *his* when searching for <-is>, for example. In a way, it is a rough counterpart to the conditions on minimal "bases" mentioned above.

The last step involved mapping all these instances of spelling variants to words as types. This leads to a crucial definition: In graphemic variation, what are types, what are tokens, and which measure is best suited to evaluate graphemic variation? In part, the answer to this question depends on the kind of linguistic unit we are interested in. In this paper, we are concerned with suffixes, and the relevant units are (graphemic) words (as opposed to letters, noun phrases, sentences etc.). On this basis, tokens are easily defined: Every individual occurrence of a word in our corpus is a token. On a more abstract level, we can then form sets of similar tokens; we will call these sets *graphemic types*. All occurrences of the 10 tokens <daungerous> in the Helsinki corpus, for example, are instances of the graphemic type <daungerous>. Because we are specifically interested in suffix variation, we can then abstract away from (put less politely, ignore) different stem spellings and subsume the respective graphemic types under *stem types*. For example, the graphemic types <dangerous> and <daungerous> are member of the stem type <dangerous>, while the graphemic types <dangerus> and <daungerus> are member of the stem type <dangerus>. The last level of abstraction is reached with the grouping of stem types to lexeme types, where the variation in the suffix is normalized. Accordingly, the stem types <dangerous> and <dangerus> are members of the lexeme type {dangerous} (the abstract morphological nature of this level is indicated by curly brackets). The relation between tokens, graphemic types, stem types, and lexeme types is illustrated in figure 1 for a small number of tokens:

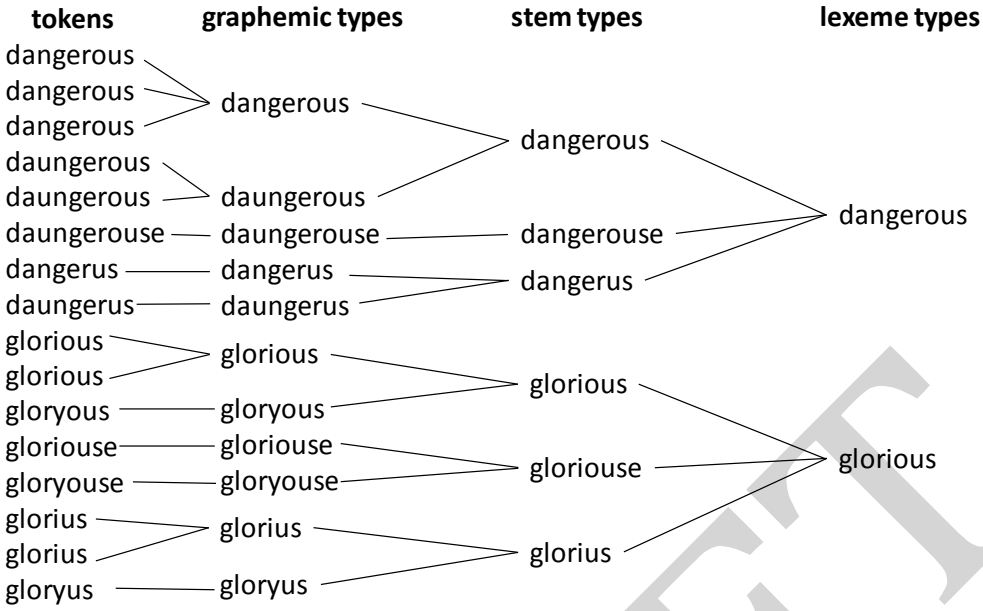

Figure 1: The relation between tokens, graphemic types, stem types, and lexeme types

This paper will focus mostly on the token counts. This does not mean that type counts are irrelevant; token counts are just easier to operationalize. However, token counts can be biased by a small number of high-frequency items that skew the data disproportionally. The usual way to handle this possible problem is to also take into account type counts. Accordingly, we will determine for each suffix spelling how many stem types there are with this suffix spelling. In the example in figure 1, for example, there are two stem types for each <ous>, <ouse>, and <us> (<dangerous>, <glorious> for <ous>; <dangerouse>, <gloriouse> for <ouse>; and <dangerus>, <glorius> for <us>, respectively).[7]

For each time span and suffix spelling, the absolute number of tokens with a given spelling is determined (e.g. all words in a given time span ending with <ous>), and the same is done for stem types (e.g. how many different stems occur with <ous>?). The ratio of token suffix spellings is determined from the absolute numbers, and the results are plotted as bar plots over time. Additionally, we provide further measures to gauge the amount of variation for each suffix. We determine how many lexemes occur with one, two, three, or four stem types in each time period. For example, in the 1420-1500 period, there are three stem types for {gracious}, <gracious>, <graciouse>, and <gracius>, but only one stem type for {religious}, namely

[7]As a consequence, the stem type numbers for the different suffix spellings do not add up to the total number of types. In the example above, there are three suffix spellings that occur with two stem types each, which is a total of six stem types - but there are only two lexeme types.

<religious> (lexemes that occur only once in a given period are excluded from this measure because by definition, they cannot show variation). On this basis, we can also calculate the mean number of stem types for each lexeme type, and the amount of lexeme types with more than one stem type.

3. Results

3.1 <ous>

Synchronically, *-ous* is a good example of how a difference in spelling mirrors a difference in morphological structure. The suffix *-ous* forms adjectives from nouns (cf. e.g. Marchand 1969: 339f.). Accordingly, the output of the word-formation rule states that phonologically, the complex word contains the stem plus [ɪs]; and the resulting word is an adjective.

There are 346 words ending in <ous> in CELEX – and they are all adjectives that end in [ɪs]. That means there is a very reliable relation between spelling and the morphological structure. Whenever readers encounter a word with final <ous>, they know it is an adjective.

Keep in mind that this may still be a reflex of phonological structure. If there were no other words that end in [ɪs], then we would basically have the same situation in phonology: All words that end in [ɪs] are adjectives; [ɪs] signals adjectivehood. Interestingly, this is precisely not the case. There are many words that end in phonological [ɪs] but not in orthographic <ous>. If we exclude the suffixes *-less*, *-ness*, *-itis*, and *-osis*, which are of the form CVC or longer, there are 666 words that end in [ɪs] in the CELEX corpus. The most prominent graphemic patterns among them (apart from ous with the above mentioned 346 words) are given in table 2:

word ending	number of words	ratio	examples
<ous>	346	52%	hazardous, nervous
<us>	117	18%	bonus, genius
<is>	72	11%	glottis, tennis
<ess>	53	8%	hostess, princess
<ice>	38	6%	office, service
rest	40	6%	

Table 2: Words in the CELEX database that end in [ɪs] (but not in <less>, <ness>, <itis>, <osis>), grouped according to their graphemic word ending

We stated above that all instances of <ous> words are potential adjectives. Of all the 320 other words in table 3, only six can be used as adjectives. If we thus cross-

classify word ending (±<ous>) and lexical category (±A), we get the following distribution:

	+<ous>	-<ous>
+A	346	6[8]
-A	0	314

Table 3: Cross-classification of word ending (±<ous>) and lexical category (±A) for all words in CELEX that end in [ɨs] (but not in <less>, <ness>, <itis>, <osis>)

We find a very stable relation in both directions: Words in <ous> are always adjectives, and the other words that end in [ɨs] almost never are; adjectives that end in [ɨs] are almost always spelled with <ous>, while non-adjectives never are. Apart from the six adjectives that are not spelled with <ous>, we have a bidirectionally unique relation between the spelling of the suffix *-ous* and the function of the suffix (formation of adjectives).

What is more, <ess> is the spelling of a separate suffix denoting female persons or animals (cf. Marchand 1969: 286ff.). Of the 53 words that end in <ess>, 47 actually refer to females (89%); there are only six exceptions (*mattress, fortress, prowess, buttress, abscess, cypress*). The relation between the spelling of this suffix and its function is also very reliable.

The English writing system makes morphology visible: You can think of <ous> as a tag attached to words that flags "adjective", while <ess> signals "noun, female person/animal" (cf. for similar phenomena in German Fuhrhop 2011). Crucially, this is information that the phonological system does not provide – it is a distinct feature of the writing system.[9]

How did this strikingly clear system evolve? To answer this question, we searched the Helsinki corpus for the following spelling variants from the OED: <ose>, <ows>,

[8] The six non-<ous> adjectives are *apprentice*, *novice*, *primus*, *bogus*, *emeritus*, and *traverse*. The status of the first three as adjectives is not clear. The only adjectival use of *apprentice* the OED cites is from 1400, while the later entries are marked as "attributive use of the singular noun". Something similar may hold for *novice*, where all cited instances could also be analyzed as nominal attributes (e.g. "targets for novice users"). *Primus* as an adjective is only attested as a postmodifier (for the elder of two persons with the same last name, e.g. "Jones primus"). These three cases are at least dubious with regard to their status as adjectives and *primus* is rare, as is *traverse* as an adjective. This leaves us with *emeritus* and *bogus*. The first is a Latin participle and the origin of the second is very unclear (see OED).

[9] As noted above, linguists use this orthographic fact as a convenient shorthand. Linguists call the English adjectival suffix [ɨs] by its orthographic form <ous>, although there are many words that end in phonological [ɨs] that do not contain the suffix, precisely because linguists have unconsciously absorbed what we have just now shown to be true: (almost) all and only words ending in <ous> are adjectives.

<is>, <owse>, <ys>, <es>[10], <ouse>, <us>, <ous>. The first way to look at the data is to determine how many different spellings of *-ous* there are for each lexeme, i.e. to determine the number of stem types. Table 4 shows the number of lexemes with one, two, three, and four stem types and the mean number of variants for all lexemes (lexemes that only appear once cannot show variation; they are grouped under the heading "- (hapax legomena)":

	1250-1350	**1350-1420**	**1420-1500**	**1500-1570**	**1570-1640**	**1640-1710**
- (hapax leg.)	5	17	30	35	40	50
1 stem type	2	10	14	25	35	49
2 stem types	2	11	5	13	5	2
3 stem types		4	8	1	1	
4 stem types		2	2			
mean number of stem types per lexeme	1.5	1.93	1.93	1.38	1.17	1.04
ratio of lexemes with more than one stem type	50%	63%	52%	36%	15%	4%

Table 4: number of types and tokens for different spelling variants of words spelled with <ous> today.

Before the 1250-1350 period, there are no useful data in the corpus. This is in line with Marchand's observation that *-ous* is an English formative from the 14th century on (Marchand 1969: 339). The height of variant spelling for this suffix is the 1350-1420 period, with 63% of lexemes having more than one stem type, and every lexeme having (on average) almost two different stem types. From then on, variation was gradually reduced, and by 1640-1710, it is marginal.

The next question is how the attested variants are distributed. Table 5 shows the number of word types and word tokens for each suffix variant, and figure 2

[10] For <-es>, searching for the word ending was not a feasible option. There are 28,488 word tokens that have at least three letters plus <-es>, which makes a manual classification far too time-consuming. The vast majority are plurals (e.g. *all my synnes*, *some thinges* etc.). Here we adopted the following strategy: After searching the corpus for the other spelling variants, we used the list of 340 graphemic types of this search (e.g. <advantageous>, <advantagious>, <affectuouse>, <ambicious>, <ambitious> etc.) as a basis for the search for <-es>-forms.

visualizes the ratio in tokens between the most frequent variants, <ous>, <ouse>, <ows>, and <us>.[11]

	1250-1350		1350-1420		1420-1500		1500-1570		1570-1640		1640-1710	
	types	tokens	types	tokens	types	tokens	types	tokens	types	tokens	types	tokens
ous	6	13	33	116	43	146	63	193	75	205	100	291
ouse	2	6	15	38	17	29	13	17	2	2		
us	2	3	9	20	13	33	10	13	2	2	1	2
ows			6	6	5	11	1	2	5	6		
os			5	5	3	3	1	1	1	2	1	1
ose	1	1	2	2	6	8					1	1
owse					1	1						

Table 5: number of types and tokens for different spelling variants of words spelled with <ous> today.

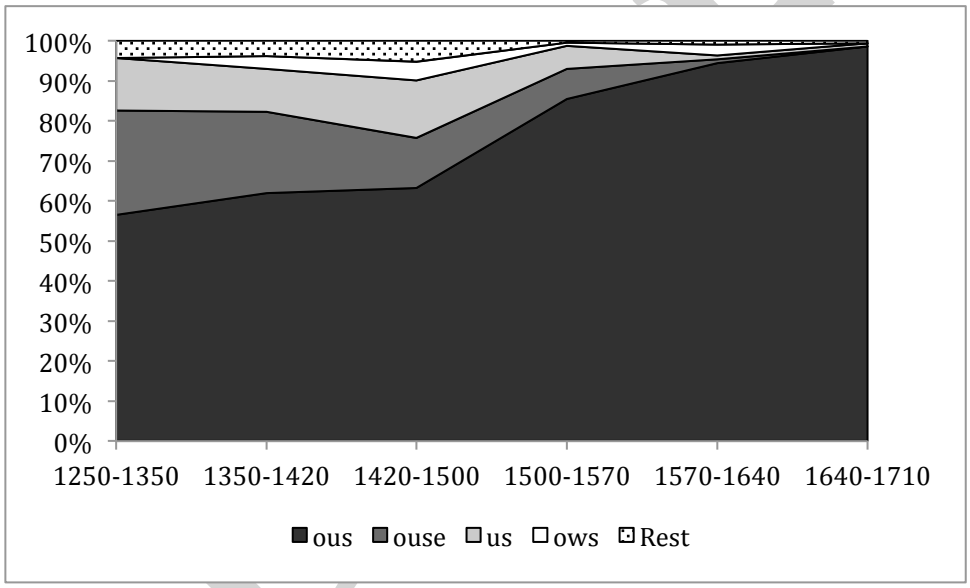

Figure 2: Relative amount of tokens in the Helsinki corpus with the suffix spellings <-ous>, <-ouse>, <-us>, <-ows>, and all other spellings. Basis: All tokens in the Helsinki corpus that end in <-ous> today.

Both types and token counts show a clear trend towards standardization, and there is no mismatch between them: ordering the spelling variants according to type

[11] Figure 2 is a line graph connecting dots that stand for time spans. A bar plot may appear to be a more truthful representation. However, the line plot has the advantage of showing how the respective amounts of each suffix are linked across the time spans. Moreover, the enclosed area for each suffix is equally big in the line plot and the bar plot.

frequency leads (with very few exceptions) to the same result as ordering them according to token frequency. Variation is reduced gradually over a period of 400 years and <ous> is the only spelling that survives. It is very unlikely that this was the effect of a conscious effort to unify the writing system. We certainly have no evidence of such a conscious effort, even among grammarians. Instead, this is an instance of a system organizing itself. One may be tempted to think that the presence of words like *bonus*, *status*, and *campus* led to the demise of the <us> variant in favour of <ous>. However, these words of Latin origin are not for the most part borrowed into English before 1550.[12] So <us> as a noun word ending became popular only after <us> as a spelling variant of –*ous* had gone.

3.2 <ic>

Like -*ous*, the suffix -*ic* is an adjectival suffix operating on nominal bases (cf. e.g. Marchand 1969: 294ff). The OED notes frequent conversions between adjectives in -*ic* and nouns (e.g. *alcoholic, arctic, classic, lunatic*; cf. OED). The word-formation rule for -*ic* states that the resulting word ends in [ɨk] and is an adjective.

There are 646 words in CELEX with final <ic>, and 628 of them are adjectives – which is a ratio of 97%. The remaining 18 words are nouns (e.g. *attic, critic, republic, logic*), some of which can also be used as verbs (e.g. *fabric, panic, traffic*). All in all, the relation between the occurrence of <ic> and the morphological structure of the respective word is very reliable: With a high probability, <ic> tells the reader that the respective word is an adjective.

Taking the phonographic perspective, there are 684 words that could potentially be spelled with final <ic>: The most frequent patterns are the following:

word ending	number of words	ratio	examples
<ic>	646	94%	allergic, demonic
<ock>	14	2%	buttock, haddock
<ick>	8	1%	derrick, rollick
<(n)ik>	5	1%	beatnik, kibbutznik
rest	11	2%	barrack, eunuch

Table 6: Words in the CELEX database that end in [ɨk], sorted according to their graphemic word ending

Apparently, there are not many words that could be spelled with <ic> but are not, compared to the great number of words that are spelled in <ic>. Still, those 38 words that end in [ɨk] but not in <ic> show a remarkable distribution: only one is an adjective (*elegiac*). The other 37 words are mostly nouns (e.g. *bannock, gimmick,*

[12] OED online lists about 1400 nouns ending in <us> and not <ous>, of which only 200 first occur before 1550.

mattock). This leads to a very clear distinction: 97% of [ɨk] words ending in <ic> are adjectives, and only 3% of those not ending in <ic> are.

	+<ic>	-<ic>
+A	628	1
-A	18	37

Table 7: Cross-classification of word ending (±<ic>) and lexical category (±Adj) for all words in CELEX that end in [ɨk]

Viewed from the perspective of the lexical category, almost every adjective that ends in [ɨk] is spelled with <ic> (except one, *elegiac*). For the non-adjectives, this correlation is not so strong: A third of them are spelled with <ic>, two thirds are not. Overall, however, we find a reliable relation between spelling and morphology. Additionally, with *-(n)ik* (e.g. *alrightnik*, *kaputnik*) we have another suffix with a distinct function and a unique spelling.[13]

To investigate the emergence of this system, we searched the Helsinki corpus for the forms <ic>, <ick>, <icke>, <ik>, <ike>, and <ique>. Table 8 shows the mean number of suffix variants that each lexeme has and the ratio of lexemes with more than one variant (relative to all lexemes with more than one occurrence).

	1350-1420	**1420-1500**	**1500-1570**	**1570-1640**	**1640-1710**
- (hapax leg.)	7	9	9	10	9
1 stem type	3	1	2	8	5
2 stem types	5	4	4	6	4
3 stem types	1		1	1	4
4 stem types		1	1	1	
mean number of stem types per lexeme	1.78	2.17	2.13	1.69	1.92
ratio of lexemes with more than one stem type	67%	83%	75%	50%	62%

Table 8: number of types and tokens for different spelling variants of words spelled with <ic> today.

There are no relevant data before 1350-1420 in the corpus. For this suffix, variation is not reduced; the last time span still features almost two suffix spellings for each

[13] This particular suffix may be less common in our sources than it is in popular language, since it tends to have a jocular connotation. However, the vowel in this suffix is probably not fully reduced, so phonology too keeps this suffix distinct.

lexeme. Standardization must have arrived after 1640-1710. The ratio of lexemes with more than one spelling peaks at the 1420-1500 period but remains relatively high afterwards. The reason for this becomes clear when we look at the type and token counts of the individual suffix variants in table 9 (figure 3 plots the ratio of tokens over time, with <y>-variants and <i> variants merged for the sake of clarity):

	1350-1420		**1420-1500**		**1500-1570**		**1570-1640**		**1640-1710**	
	types	tokens	types	tokens	types	tokens	types	tokens	types	tokens
ic	5	10	1	1					7	18
ick			1	1	2	3	6	8	20	76
icke					3	5	15	29	1	1
ik	9	28	3	9			1	1		
ike	2	3	2	2	11	29	5	6		
ique	1	2	6	6	4	9	10	30	6	15
ycke			1	1	1	1				
yk	4	4	5	5	1	2				
yke	2	2	3	4	4	9				

Table 9: number of types and tokens for different spelling variants of words spelled with <ic> today.

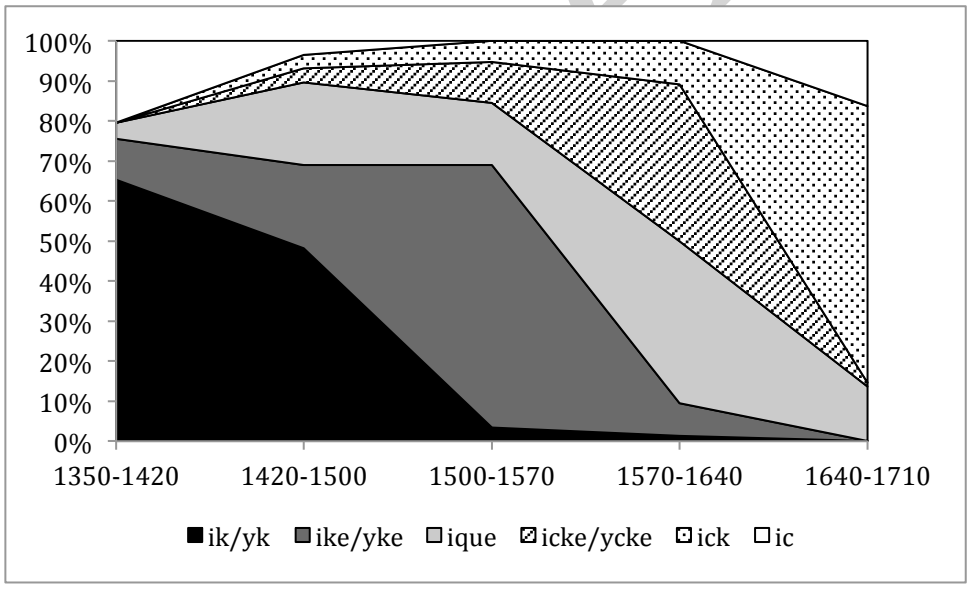

Figure 3. Relative amount of tokens in the Helsinki corpus with the suffix spellings <-ick>, <-ik/-yk>, <-icke/-ycke>, <-ike/-yke>, <-ique>, and <-ic>. Basis: All tokens in the Helsinki corpus that end in <-ic> today.

Compared to the development of *-ous*, there is much less uniform movement in the data. Variation is not gradually reduced (like it was for *-ous*); on the contrary, new

variants are introduced in different time spans: <ick> appears 1420-1500; <icke> is first attested 1500-1570, and <ic>, today's spelling, resurfaces just 1640-1710 after two dormant time spans. These variant spellings take the lead one after the other: For 1350-1420, <ik> is the major variant; then it gradually fades out. For the period 1420-1500, <ique> takes over (at least for the types); for 1500-1570, <ike> is the most popular spelling; and for 1570-1640, it is <icke>. In the last time span in the Helsinki corpus, <ick> is dominant.

So, contrary to one variant gradually gaining strength and the other variants diminishing, we find a quick succession of variants that emerge gradually, peak, and then decline. The sequence is interesting and will be discussed in the conclusion.

(1) <ik> → <ique> → <ike> → <icke> → <ick>

Today's spelling, <ic>, is a minor variant in the last period. To find out when the shift from <ick> to <ic> occurred, and how quick it was, we used the Google Books corpus, British section via the interface americancorpus.org.[14] To make the data comparable, we searched for all the 39 lexemes in <ic> that occurred in the two last periods in the Helsinki corpus. We concentrated on the <ic>, <ique>, and <ick> variant of each word, leaving the minor variants aside. Of course, Google Books is a much larger corpus than the Helsinki corpus. In the last time span in the Helsinki corpus, there are 110 tokens; in the next 70 years of the Google Books corpus, there are almost 300,000 tokens. This is not surprising, considering that the Helsinki corpus has about 1.5 million words, while the British sub-corpus of Google Books encompasses 34 billion words. However, the two data types seem to be roughly connectable (compare the last time period in figure 3 and the first in figure 4). Because the temporal resolution that americancorpus.org provides is much higher, we get a finer-grained graph. Figure 4 picks up where figure 3 above leaves off:

[14] This interface is convenient because, while it operates on the Google Books data just like Google's own NGram viewer, its output are actual frequencies (not just graphs, or ratios), which can then be added for each of the words that were searched. One disadvantage is that initial minuscule and majuscule spellings have to be searched separately and combined in a later step.

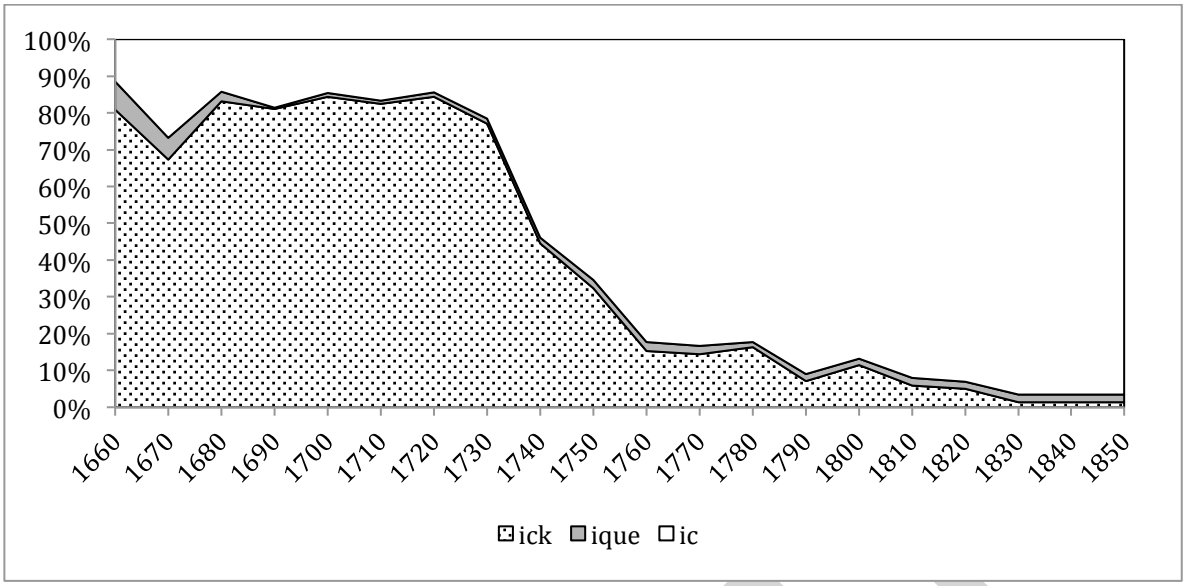

Figure 4: Ratio of the token count of 34 word types in three spelling variants, <ic>, <ick> and <ique>. Data source: Google Books, British sub-corpus, via americancorpus.org.

The variant <ique> is marginal throughout the covered time spans, yet it does not fully vanish. The respective words could actually be French words cited in Google Books. The variant <ick> gradually declines, and the crossover point by which <ic> becomes dominant is the decade 1740-1750. After that, <ick> is still attested, but it is very slowly washed out of the system, with <ic> the only remaining variant.

3.3 <al>

Two suffixes *-al* can be distinguished. One is nominal *-al* (type *arrival*, cf. Marchand 1969: 236f.), and the other is the more frequent adjectival *-al* (type *accidental*, cf. Marchand 1969: 238ff.).[15] Accordingly, there are two word-formation rules, and the outputs state that the resulting word should end in [ɨl] and be a noun, or end in [ɨl] and be an adjective. This homophony of suffixes is obviously not resolved in spelling; the two suffixes are homographic. Additionally, we find frequent conversions between nouns and adjectives (e.g. *capital, final, vocal*). From this perspective, it makes sense to determine in how many cases words with *-al* are either nouns or adjectives (as opposed to other parts of speech).

There are 913 words in CELEX that end phonologically in [ɨl] and graphemically in <al>. Of these, only 22 can be used as verbs (e.g. *equal, local, metal, pedal, rival, spiral*). The majority of these 22 words are conversions of other word categories;

[15] Etymologically, Latin had an adjectival suffix *–ālis*, which could also be used in the neuter plural form *–ālia* as a noun. In French, the two became quite distinct as *–el* and *–aille* but the two eventually collapsed again in form in borrowings into English.

only three words can exclusively be verbs (*outgeneral, outrival, victual*).[16] Taking the most liberal stance (the worst case scenario, so to speak), there are 22 out of 913 words where spelling does not indicate lexical category (2%). In other words, we again find a reliable relation between spelling and morphology. Note that this relation does not involve a unique mapping of spelling and suffix: *-al*$_1$ and *-al*$_2$ are not differentiated. Rather, what is indicated in spelling is that the graphemic word in question bears either one of the two suffixes – or that it is probably not a verb.

The 913 words mentioned above are not the only ones that could potentially be spelled with final <al>; in total, there are 1,511 words that phonologically end with [ɨl] (excluding words that end in -able/-ible and -ful, where [ɨl] is part of a bigger suffix). Table 10 lists the most frequent patterns:

word ending	number of words	ratio	examples
<al>	913	60%	liberal, regimental
<le>	410	27%	crumble, thistle
<el>	126	8%	channel, shovel
<il>	32	2%	devil, pencil
<yl>	9	1%	ethyl, vinyl
rest	21	2%	gambol, pistol

Table 10: Words in the CELEX database that end in [ɨl], sorted according to their graphemic word ending

As is obvious from table 10, a considerable number of words could be spelled with <al>, but are not. At least one of the spelling variants is a suffix in its own right, <yl>; it denotes chemical radicals (e.g. *acetyl, ethyl, methyl, vinyl*). <el> and <le> are spelling variants of a suffix that is no longer productive (cf. OED). Marchand (1969) distinguishes two types with distinct histories, iterative verbs of the type *sparkle* (Marchand 1969: 322f.) and mostly diminutive nouns of the type *spittle* (Marchand 1969: 324).

The OED states that <le> is the default variant, and that <el> appears "after ch, g soft, n, r, sh, th, and v" (OED). Apparently, the two spellings are in complementary distribution. According to CELEX, there seem to be additional environments that trigger <el>. With minor exceptions, <el> appears after the single letters <m, n, r, s, v, w> and after the combinations <ch, sh, th>. The alternative <le>, on the other hand, appears after the single letters <b, c, d, f, g, k, p, t, x, z> and after <ck>. Moreover, if <c> and <g> correspond to continuants (as in *cancel* or *angel*), they are

[16] The last one can clearly be used as a noun (cf. e.g. OED) – it just not annotated as such in CELEX. However, we will stick to the methodology sketched out above and count all three as verbs. Otherwise, we would have to check every lexical category in CELEX against the OED.

always followed by <el>. This last fact is not surprising: in these cases <e> functions as a marker of the "soft" fricative correspondence (cf. Venezky 1999: 84).

From the discussion of the previous two suffixes one might expect a complimentary distribution of word classes: Words with <al> are adjectives and nouns, words with other endings are neither adjectives nor nouns. The picture for *-al* is more complex, however. Only 15 words that do not end in <al> are adjectives (and cannot be used as nouns or verbs), e.g. *ample*, *civil*, *feeble*. But nouns are quite common, and there are frequent conversions between nouns and verbs (e.g. *model*, *quarrel*, *trouble*). This situation is the mirror image of the <al>-words: There we had frequent conversions between nouns and adjectives (*capital, final, vocal*), and almost no verbs were spelled with <al>. For the non-<al>-words, we have frequent conversions between nouns and verbs, and almost no adjectives are spelled that way. Table 11 shows the distribution:

	+<al>	-<al>
+A	910	15
+N		583
+V	3	

Table 11: Cross-classification of word ending (±<al>) and lexical category (A, N, V) for all words in CELEX that end in [ɨl]

Even though the distribution is more complex than for the last two suffixes, the relation between spelling and morphology is reliable: Words with <al> are nouns or adjectives, but not verbs, and words with other endings (mostly <le> and <el>) are nouns or verbs, but not adjectives.

Diachronically, not only can we track the developments of <al>, but there is also sufficient data to investigate the development of <le> and <el>. Starting with <al>, we searched the Helsinki corpus for the forms <ale>, <alle>, <ell>, <el>, <al>, <all>. Table 12 shows how much variation there is in each time period:

number of variants	**1350-1420**	**1420-1500**	**1500-1570**	**1570-1640**	**1640-1710**
- (hapax leg.)	18	25	25	38	58
1 stem type	15	6	11	37	36
2 stem types	10	13	27	20	28
3 stem types	4	7	5	1	
4 stem types	2	2	1		
5 stem types		1			
mean number of stem types per lexeme	1.77	2.28	1.91	1.38	1.44
ratio of lexemes with more than one stem type	52%	79%	75%	35%	43%

Table 12: number of types and tokens for different spelling variants of words spelled with <al> today.

	1350-1420		**1420-1500**		**1500-1570**		**1570-1640**		**1640-1710**	
	types	tokens	types	tokens	types	tokens	types	tokens	types	tokens
all	8	21	32	145	53	305	85	422	38	119
al	32	168	27	71	47	141	30	70	111	456
el	12	37	6	14			1	2	1	2
alle	4	4	14	31	2	2	1	1		
ale	10	16	4	5	6	12	1	1		
ell	3	4	6	13	3	3	1	1		
ille	1	6								
ayle	1	1	1	2	1	1				
le	1	2								
yle	1	1								
aile			1	1						

Table 13: number of types and tokens for different spelling variants of words spelled with <al> today.

Variation peaks in the 1420-1500 period and gradually declines afterwards (the slight rise in the last period is due to the late emergence of today's form; see figure 5 below). Until the 1500-1570 period, variation is the rule rather than the exception for lexemes with this suffix. This is mirrored in table 13, above, which shows the distribution of the variants over time, for both types and tokens, and figure 5, which

plots the relative number of tokens over time (the minor variants are grouped together).

Figure 5: Relative amount of tokens in the Helsinki corpus with the suffix spellings <-all>, <-al>, <-alle>, <-el>, and all other spelling variants. Basis: All tokens in the Helsinki corpus that end in <-al> today.

In 1350-1420, today's form, <al>, is the most frequent spelling. The relative amount of <all> rises steadily until the 1570-1640 period (where it reaches >80%), and then quickly declines. By 1640-1710, <al> is the dominant form again. This final transition is rather quick. From 1500-1570 on, there are only two major variants, <all> and <al>. Note again that the succession of suffix spellings is in some ways similar to the one found for *-ic* (cf. section 4.1 for further discussion).

(2) <al> → <all> → <alle> → <al>

For <el>, we searched the Helsinki corpus for the forms <el>, <le>, <ell>, <elle>, <ele>. As the main focus is on the evolution of <al>, we omit the discussion of overall variation for this word ending and only present the bar plot (figure 6), which tracks the distribution of tokens over time:

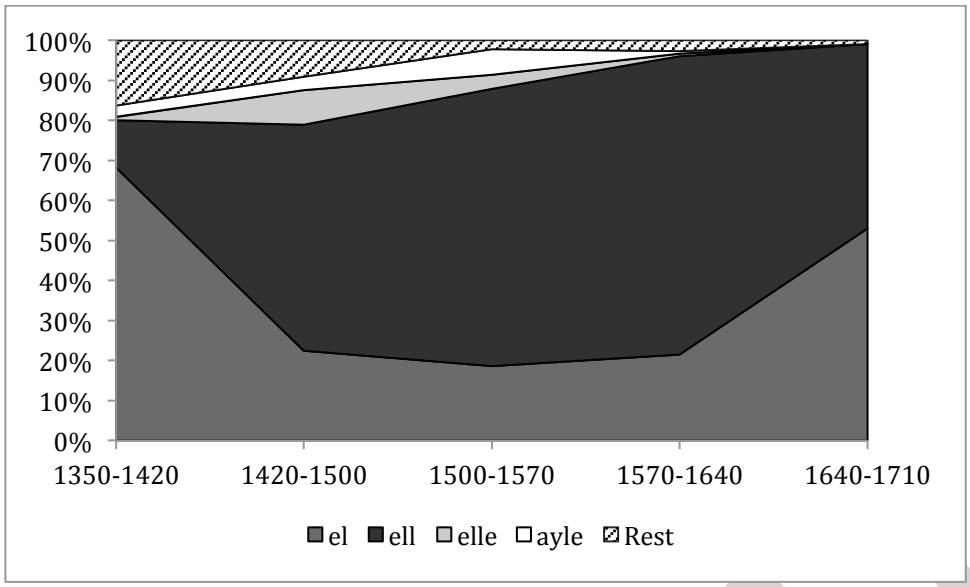

Figure 6: Relative amount of tokens in the Helsinki corpus with the suffix spellings <-ell>, <-el>, <-elle>, <-ayle>, and all other spelling variants. Basis: All tokens in the Helsinki corpus that end in <-el> today (1350-1420: N=110; 1420-1500: N=241; 1500-1570: N=223; 1570-1640: N=153; 1640-1710: N=100).

The number of attested variants is much bigger for <el> than for the other suffixes we have investigated so far, but there are only two major variants, <el> and <ell>. Similar to <al>, today's spelling was the dominant one 1350-1420, but the variant with the doubled final consonant <ell> steadily rises until 1570-1640; in the following time period, <el> is again the major variant, although variation is far from resolved. Note again the similar pattern of spelling variants:

(3) <el> → <ell>, <elle> → <el>

For the last suffix, <le>, we searched the Helsinki corpus for the forms <ale>, <alle>, <ell>, <el>, <al>, <all>. Again, we only present the bar plot (figure 7), which plots the ratio of tokens with a given suffix variant over time:

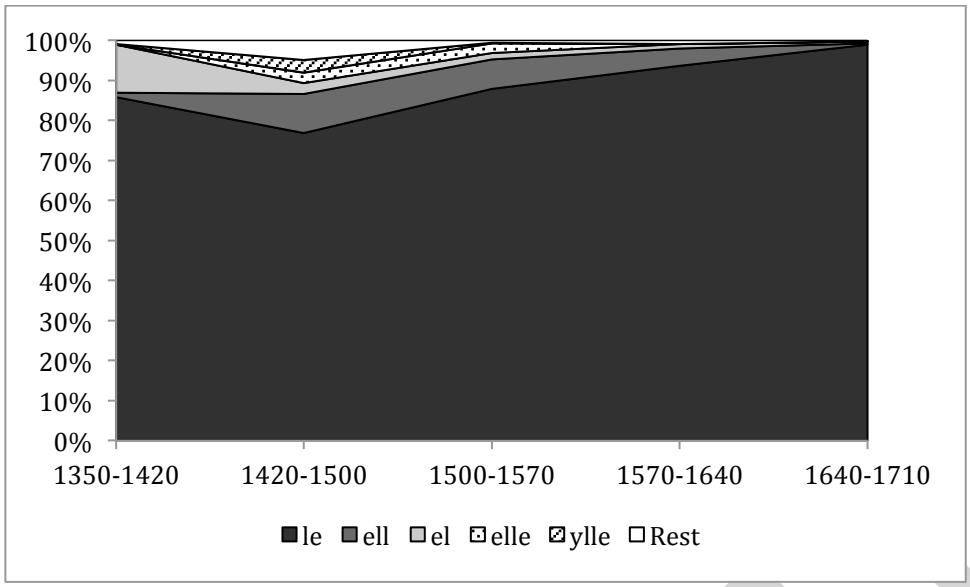

Figure 7: Relative amount of tokens in the Helsinki corpus with the suffix spellings <-le>, <-ell>, <-el>, <-elle>, <-ylle>, and all other spelling variants. Basis: All tokens in the Helsinki corpus that end in <-le> today (1350-1420: N=865; 1420-1500: N=794; 1500-1570: N=1021; 1570-1640: N=999; 1640-1710: N=1024).

For <le>, as many variants are attested as for <el>, and again most of them are marginal. Today's spelling, <le>, was dominant throughout all five time periods, and it continued to become more frequent from 1420-1500 until 1640-1710. The other spelling variants were slowly washed out of the system.

3.4 <y>

There are several homophonous suffixes *-y*. Depending on the point of view, the precise number varies. The Oxford English Dictionary presents a list of six *-y* suffixes, mostly on an etymological basis. Bauer, Lieber & Plag (2013), on the other hand, posit three *-y* suffixes on synchronic grounds, and we will follow their classification here:

- Adjectival *-y* (type *windy*, *choosy*): This suffix (on nominal and verbal bases) is 'very productive' (Marchand 1969: 352f).
- Nominal *-y* (type *harmony*, *family*): This suffix is 'somewhat elusive', as Bauer, Lieber & Plag (2013: 255) state; for many nouns with final <-y> it is unclear whether they actually are morphologically complex (e.g. *family*).
- A (nominal) diminutive suffix *-y* (type *granny*), which has <-ie> as a spelling variant.

Accordingly, there are three word-formation rules, and the outputs state that the resulting word should end in [i] and be a noun, or end in [i] and be a diminutive noun, or end in [i] and be an adjective. This homophony of suffixes is not resolved in spelling; all suffixes are homographic.

There are 1,249 words in the CELEX corpus that end in [i] and <y> (excluding <y> as part of the larger suffixes *-acy*, *-ary/ery/ory*, *-ancy/ency*, *-ey*, *-ity*, *-ly* and *-ry*). 134 of them can be used as verbs, among other lexical categories (e.g. *belly, candy, lobby, ready*). For many of them, a verbal use seems to be the exception, and only 11 out of those 134 words are exclusively verbal and cannot be used as nouns or adjectives (e.g. *accompany, bury, embody, marry*).[17] Again taking the most liberal stance, we end up with 133 potential verbs out of 1,249 words, or 11%. Another way to put it is that 89% of words that end in <y> cannot be used as verbs. In other words, the relation between spelling and morphology is quite stable, though there are some exceptions.

Looking at words that could potentially be spelled with final <y> but which are not, we find that there are 1,511 words in CELEX. Table 14 lists the most prominent spellings for the phonological ending [i]:

word ending	number of words	ratio	examples
<y>	1,249	83%	dreamy, harmony
<i>	74	5%	Israeli, spaghetti
<ey>[18]	65	4%	chimney, money
<ie>	63	4%	brownie, sweetie
<e>	49	3%	recipe, karate
rest	11	1%	chassis, coffee

Table 14: Words in the CELEX database that end in [i], sorted according to their graphemic word ending

So 89% of words with final <y> are nouns or adjectives, and there are a number of words that could potentially be spelled with final <y>. What about the lexical category of these words? Does the spelling convey information about lexical category? Correlating lexical category (A/N vs. V) and word ending (±<y>) leads to the following results:

[17] It is striking that prefixation is involved in 5 of these 11 verbs (*remarry*, *intermarry*, *embody*, *disembody*, *miscarry*). Final <y> does not seem to be a good ending for English verbs.

[18] Twelve words with final <ey> were analyzed as the suffix *-y* on bases ending in <e> (e.g. *dopey*, *homey*, *pricey*). They appear in the <y>-category of the table.

	+<y>	-<y>
+A	1116	34
+N		128
+V	133	

Table 15: Cross-classification of word ending (±<y>) and lexical category (A, N, V) for all words in CELEX which end in [i]

As with <al>, the presence of <y> has a negative value: It signals that the word is probably not an English verb. On the other hand, the absence of <y> (and the presence of other word endings that correspond to [i]) signals that the word is probably not an adjective. The figures in table 15 become even clearer when we look at the 34 adjectives that do not end with <y>. Half of them are words ending in <i> (e.g. *Afghani, Israeli, Qatari*, see below). We would not want to call these words exceptions - they bear a suffix that can regularly form nouns that can also be used as adjectives. Obviously, the word endings in table 15 do convey some kind of morphological information, but this information is more complex than just lexical category membership. This becomes clear when we look at the two endings that are productive suffixes in their own right, <i> and <ie>.

The suffix *–i* is an ethnonym (cf. Marchand 1969a: 354f.;, Marchand 1969b). It denotes people from Eastern or Near-Eastern countries. Only 22 of the 74 words with <i> bear this suffix (e.g. *Israeli, Bahraini, Kuwaiti, Pakistani*), the rest are mostly foreign words of Italian (*vermicelli, broccoli, ravioli, salami*) or Hindi origin (*sari, rani, kukri*). Yet *-i* is very regular and productive in this limited domain as an ethnonym. In combination with the capitalization, words like *Pakistani* indeed indicate a certain morphological function – and what is just as important, a phonographically possible spelling like *<Pakistany> is not an option.

The suffix *-ie* was introduced above as a spelling variant of diminutive *-y* (cf. OED, Marchand 1969:298f.). Of all the 63 <ie> words in CELEX, 40 are classified as diminutives or hypocorisms in the OED (63%). The other 37% are words of mostly French origin (e.g. *brasserie, gendarmerie, patisserie*). What is more, this suffix is clearly productive. A very tentative search of the Corpus of American English (CoCA, http:// http://corpus.byu.edu/coca/) reveals a great number of non-lexicalized ad-hoc formations, as the following randomly chosen examples show:

(4) a. I must have looked puzzled, because she explained that a Cliffie is someone who goes to Radcliffe

b. In this article, Marilyn shows a "flattie" (that's stereo photographer language for a non-stereo picture!)

c. I can see the headlines: narco, trannie, and journalist crash on way TO FOREST OF THE WHORES (sic)

Intuitively, the spellings <Cliffy>, <flatty> and <tranny> do not work as well in the given contexts; <ie> seems to be the preferred variant in these cases. Obviously, a lexical database like CELEX cannot capture these formations; we must conclude that they are more frequent than the ratio in CELEX (63%) suggests.

Additionally to *-i* and *-ie*, there is a third productive suffix that can be pronounced [i], *-ee*. It is mostly stressed (Bauer, Lieber & Plag 2013), but in some cases, the stress shift is reversed, leading to unstressed final [i], e.g. in *employee*, which can be pronounced /əmˈplɔ(ɪ)ˌi/. In this respect, <ee> is a further case that could potentially be spelled <y>, but where the spelling indicates a certain special function.

Finally, <ey> must also be taken into account. Apart from the predictable cases mentioned above (fn. 18), the words it occurs in are mostly unanalyzable (e.g. *chimney, chutney, volley*). There seems to be no unique distribution of <ey> (as opposed to <y>), but there are two noteworthy observations. Firstly, <ey> occurs mostly after <l> or <n> (e.g. *trolley, parsley*; *jitney, honey*). However, this position is not specific to <ey> (cf. e.g. *early, jolly*; *pony, tiny*). Secondly, words that end with <ey> are only rarely adjectives. Of the 65 words with final <ey>, ten can be used as adjectives (e.g. *medley, motley, phoney*). In three of these ten cases, (*clayey, gooey, phooey*) <ey> is clearly isomorphic to adjectival <y>. This distribution can be captured with a graphemic rule that bans <y> after stem-final vowel clusters and demands <ey> instead. Summing up, <ey> may be interpreted as a non-adjectival spelling of [i].[19]

Diachronically, only the search for words spelled with final <-y> today yielded enough results; there is not enough data for the potentially interesting development of today's *-ie*, *-i*, and *-ey*. We searched the Helsinki corpus for the following spelling variants of today's *-y* suffix: <y>, <ie>, <ey>, <ye>, <i>, <ee> and <e>.[20] We excluded, as in the synchronic investigation, the other existing suffixes *-acy*, *-ary/ery/ory*, *-ancy/ency*, *-ey*, *-ity*, *-ly* and *-ry*. Table 16 shows an overview of the amount of

[19] The OED formulates a similar rule: "When the suffix is appended to a n. ending in *y*, the convention of modern spelling requires it to be spelt *-ey*". Note that this rule does not capture the spellings <gooey> and <phooey>. It does cover cases like <skyey, *skyy>, but these cases can also be captured with a constraint that prevents most vowel letters from doubling.

[20] For the variant <e>, we had to adopt a different search strategy. Simply searching for all words with final <e> in the respective time spans of the Helsinki corpus leads to more than 130,000 hits, far more than we could reasonably filter manually. Instead, we opted for the following strategy: After searching the corpus for the six other spelling variants, we used the list of 782 graphemic types of this search (e.g. <agony>, <albany>, <allemyghtty>, <allemyghty>, <allmightie>, <allmyghty> etc.) as a basis for the search for <e>-forms.

variation per time span (before the 1250-1350 time span, the data were too sparse to be useful):

number of variants	**1250-1350**	**1350-1420**	**1420-1500**	**1500-1570**	**1570-1640**	**1640-1710**
- (hapax leg.)	26	54	65	59	89	119
1 stem type	20	41	48	32	63	155
2 stem types	23	32	47	61	84	25
3 stem types	6	14	22	49	32	
4 stem types	4	9	7	10	5	
5 stem types			1	2		
6 stem types				1		
mean number of stem types per lexeme	1.89	1.86	1.93	2.30	1.89	1.14
ratio of lexemes with more than one stem type	62%	57%	62%	79%	66%	14%

Table 16: number of types and tokens for different spelling variants of words spelled with <y> today.

The amount of variation is roughly the same for the first three time periods. It then peaks in the 1500-1570 period, before it almost fully declines over the next two periods. Table 17 shows the distribution of the variants over time, for both types and tokens, and figure 8 plots the ratio of tokens with a given suffix variant over time:

	1250-1350		1350-1420		1420-1500		1500-1570		1570-1640		1640-1710	
	types	tokens	types	tokens	types	tokens	types	tokens	types	tokens	types	tokens
y	31	220	74	821	141	1684	156	1303	210	1764	288	2941
ie	36	168	60	250	56	140	137	573	174	1013	30	72
ye	15	29	42	231	71	202	97	350	42	80	2	2
i	33	343	31	192	19	43	9	11	4	4		
e	9	26	11	97	12	108	7	40	3	5		
ee	2	2	16	143	4	14	4	9				
ey			3	77	3	11	6	33	3	34	4	7

Table 17: number of types and tokens for different spelling variants of words spelled with <y> today.

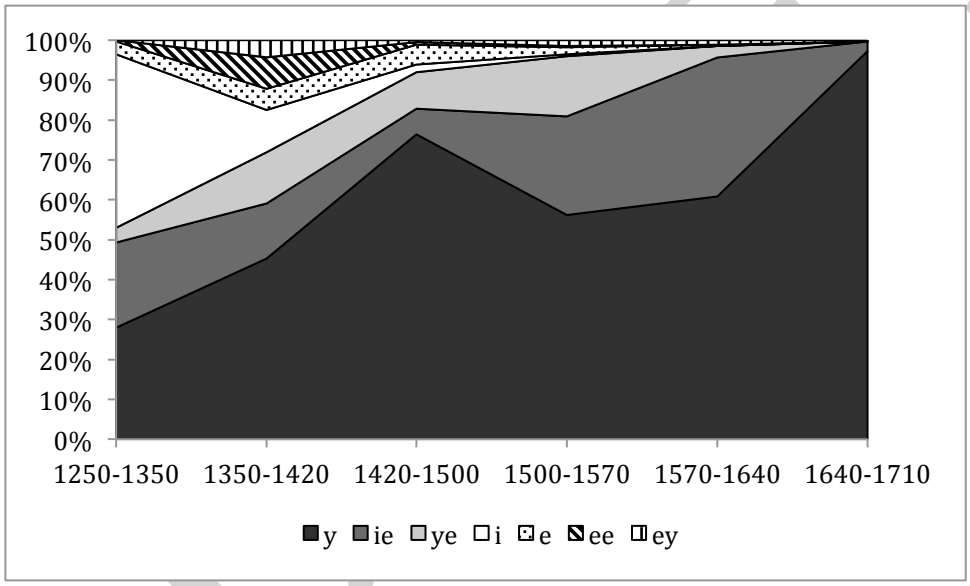

Figure 8: Relative amount of tokens in the Helsinki corpus with the suffix spellings <-y>, <-ie>, <-ye>, <-i>, <-e>, <-ee>, and <-ey>. Basis: All tokens in the Helsinki corpus which end in <-y> today.

Based on the token count, we can see that the most frequent spelling in the 1250-1350 period is <i>, which then quickly diminishes; by 1500-1570, it is gone. We can also observe an early tendency towards standardization until the 1420-1500 period: <y> was gradually becoming the most frequent variant. The next period, however, sees two rather minor variants in 1420-1500 gain weight, <ye> and <ie>. The proportion of <y> spellings declines to just above 50%. Interestingly, with the exception of <ye>, all spellings are actually still in use today, but mostly for different functions. As shown above, <i> is used to mark ethnonyms, <ee> marks patient formations, and <ie> marks diminutives.

This functionalization of leftovers supports the competition-based approach introduced earlier: In processes of standardization, only those spelling variants survive that can find distributional or functional niches. This is clearly the case for <ie>: As a spelling for today's adjectival or nominal *-y*, <ie> was marginal by the 1640-1710 period (cf. figure 8). As a spelling for the diminutive suffix, the OED cites as the earliest instances 1595 *dummie*, 1663 *grannie*, 1681 *dearie* and 1693 *mousie*. It seems that <ie> was able to take on its new role as a spelling for the diminutive suffix at the very time it was discarded as a spelling for adjectival and nominal *-y*.

For <i>, much more time passed between its disuse as a variant for today's nominal or adjectival *-y* and its new function as the spelling for the ethnonym suffix. <i> was discarded as a spelling for today's *-y* by the 1500-1570 period at the latest (cf. figure 8), but the ethnonym suffix only became productive in the 19th century (cf. OED *-i* suffix[2]). Still, the fact that <i> was no longer in use as a spelling for today's *-y* at that time (e.g. *<windi>, <family>) must have been beneficial for the prevalence of the ethnonym spelling <i>.

On this basis, we can make a prediction about the future development of diminutive *-y*/*-ie*. Today, we find formal variation between both spellings (see above). But <y> is also the standard spelling for adjectival and non-diminutive, nominal *-y*, that is, we find functional variation for the spelling <-y> (both as a diminutive and a non-diminutive suffix). On this basis, we predict that eventually <ie> will become the dominant spelling for the diminutive suffix. This is not to say that all functional variation in the system will be resolved at some point. But this case is special because a distinct spelling for diminutives actually exists as a variant, and as we have shown throughout this paper, the English writing system tends to mark morphology in spelling in comparable cases.

4. Conclusion

For each suffix we investigated, we found that the spelling marks morphological information in some way or other. Homography of suffixes and homophonous word endings is avoided in the majority of cases.

- <ous> signals that the word is an adjective. There is a very clear, almost complementary distribution. If a word ends in <ous>, it is an adjective (e.g. <nervous>; if it does not end in <ous> (but phonologically in [ɨs]), it is not an adjective (e.g. <service>).
- <ic> also marks words as adjectives, although the distribution is not as unequivocal. If a word ends in <ic>, it is almost always an adjective (e.g. <sonic>); if it does not end in <ic> (but phonologically in [ɨk]), it is (with one exception) not an adjective (e.g. <gimmick>)
- <al> signals that the word is an adjective or a noun; at the same time, there are frequent conversions between them (e.g. <capital>). Words that do not

end in <al> (but phonologically end in [ɨl]), on the other hand, are either nouns or verbs (but not adjectives) – and again, we find frequent conversions between these categories (e.g. <model>).

- <y> signals that the word is probably not a verb; words that do not end in <y> (but phonologically end in [i]) are probably not adjectives. Words that end in <i> are ethnonyms or foreign borrowings. Words that end in <ie> mark diminutives.

These features are unique to spelling; in phonology, this information is not encoded: the spelling distinguishes homophonous suffixes or word endings. This can be argued to serve the needs of silent reading: tagging words as adjectives, verbs or nouns potentially enables readers to build up syntactic structure and access information more quickly.

Looking back, we traced the evolution of this system back in time for close to a millennium. For most suffixes, there was a considerable amount of variation in spelling at first, but variation was gradually washed out of the system. In the following, we will discuss two interesting facets of this development: the patterning in the succession of spelling variants (4.1) and co-variants of this variation (4.2.).

4.1 Patterning of variants

For three suffixes, we observed a characteristic succession of spelling variants. They are reproduced here as (5.a-c):

(5) a. <ik> → <ique> → <ike> → <icke> → <ick> → <ic>
b. <al> → <all> → <alle> → <al>
c. <el> → <ell>, <elle> → <el>Common to all three suffixes is the employment of two different means, consonant doubling and final <e>. For <al> and <el>, final <e> does not occur without consonant doubling (there is no form *<ale>, *<ale>). Apart from this, the suffixes show a remarkable uniformity in their development: They start off as simple <VC>-structures, then go through stages of <VCC> and <VCCe> until finally returning to <VC>. Apparently, the concrete forms a given suffix appears in are not haphazard, but follow general trends in the writing system. The development in the spelling of suffixes is part of the general trend that the spelling of English words is subject to.

At least for the double consonant variants, a possible reason for their demise lies in the marking of prosodic structure. The effect of consonant doubling is visible in words from French: from ca. 1500, French words changed their spelling from single intervocalic consonant in French to double consonant in English, e.g. OFr. <bagage> > ME <baggage>, which mirrors the change in foot structure from iamb to trochee [σ'σ] > ['σσ] (cf. Upward/Davidson 2011: 179). Under this view – i.e. if double consonants are employed to mark the preceding syllable as stressed – spellings like <demonick> are dysfunctional because they imply a stressed ultima.

However, the concrete temporal progression of forms does differ: For example, the <VCC> form for today's *-ic* peaks 1640-1710 and then gradually declines, while the analogous forms for *-el* and *-al* peak 1570-1640. Likewise, <VCCe> forms for today's *-ic* peak later than those for *-el* and *-al*. (1570-1640 vs. 1420-1500). This speaks against the hypothesis that double consonants and final <e> were used by printers to achieve an even outer margin (cf. Scragg 1974: 71f.): In that case, we would expect no effect of the actual suffix. Thus, while some general spelling principles seem to be at work guiding the inventory of possible forms, the unification of spelling – the emergence of culture out of anarchy – is different for each suffix.

4.2 Co-variants of diachronic variation

An interesting question to ask is whether this diachronic variation co-varies with other factors. One such factor could be lexemes. It is conceivable that the total amount of variation stems from different lexemes with distinct (but consistent) spellings of the suffixes. If, for example, *capital* was always spelled <capitall>, and *natural* always <natural>, on the whole we would find variation between <-all> and <-al>, but this variation would be entirely explicable with reference to words. This example may seem a little far-fetched, but lexically based linguistic change has been widely discussed for over a century (Labov 1994).

The other possibility is that diachronic variation co-varies not with lexemes, but with texts. It is conceivable that texts themselves are consistent, and the overall variation arises from the investigation of many texts with different standards. For example, if *-al* is always spelled <-al> in one text, but <-all> in another, on the whole we would find variation between <-al> and <-all>.

In a nutshell, the question we will follow now is: What is more consistent, words or texts? To answer it, we take a closer look at one suffix, *-al*, and use a statistical measure for dispersion. Wilcox's (1967) VarNC is a measure to determine variance in nominal distributions. It ranges between 0 (all instances are in one group, e.g. all spellings of today's *-al* are <all>) and 1 (all instances are equally distributed over the groups, e.g. each of the possible *-al*-spellings has the same number of occurrences). If a spelling for *-al* only occurs once in a given text, or is attested only once with a given lexeme (within one time period), it is excluded. This way, we only investigate lexemes and texts that can potentially vary.

For each time period, we calculated the following three measures:

- VarNC overall: this measure is calculated over the total token counts for each period (cf. table 13 above).
- VarNC lexemes: for this measure, a list of lexemes is generated together with the absolute token counts for each suffix spelling. For example, *special* occurs in the 1420-1500 period with the spellings <-al> (18x), <-all> (23x), and <-

alle> (9x). For each lexeme, VarNC is calculated; the value presented in table 22 below is the mean of all VarNC values for the given time period.

- VarNC texts: analog to the lexemes, a list of texts is generated together with the absolute token counts for each suffix spelling in these texts. For example, in "The Trials of Sir Nicholas Throckmorton" (1500-1570 period), <-al> occurs 33 times, <-all> 30 times.

	varNC overall	varNC lexemes	varNC texts
1350-1420	0.60	0.23	0.43
1420-1500	0.72	0.42	0.27
1500-1570	0.52	0.36	0.27
1570-1640	0.28	0.13	0.13
1640-1710	0.37	0.19	0.17

Table 18: Wilcox's VarNC for the overall token count of -al-words (varNC overall); varNC for tokens grouped according to lexeme type (varNC lexemes); varNC for tokens grouped according to text (varNC texts).

The results for the overall variation are in line with what we presented above using a simpler measure (cf. table 13): Variation peaks in the 1420-1500 period and then gradually declines, with a slight rise in the latest period.

Except for the first time period (1350-1420), variation within texts is lower than variation within lexemes (or equally low, as in the 1570-1640 period). That means texts are more consistent than lexemes, and the pursuit for consistency was earlier for texts than for lexemes.

4.3 General conclusion

The goal of this work is to present in some empirical depth two related findings about the spelling of English derivational affixes. The first and simpler finding is that a number of derivational suffixes in contemporary written English have remarkably regular distinct spellings that differentiate them consistently from homophonous strings, some of which are used for other suffixes. This finding extends previous research showing that English spelling is both lexical and morphological to a greater extent than other alphabetical writing systems (Bolinger 1946, Chomsky & Halle 1968, Venezky 1999: 197ff.). The second finding is much broader. We have investigated the history of these systematic regularities in suffix

spelling through almost a millennium of the written language. Throughout this entire period, there is no evidence of an external authority having any influence over the spelling of these suffixes.[21] Instead, the regular spellings emerged gradually, through a sorting out process of competition between alternate spellings. To the extent that regular spelling has become an integral part of the culture of the English language, it has truly emerged out of anarchy.

REFERENCES

Anderson, Stephen R. 1992. A-morphous Morphology. Cambridge: Cambridge University Press.

Aronoff, Mark. 1976. Word formation in generative grammar (Linguistic Inquiry Monographs No. 1). Cambridge, Mass.: MIT Press.

Aronoff, Mark. 1978. An English spelling convention. Linguistic Inquiry 9.299-303.

Aronoff, Mark. 1985. Orthography and linguistic theory: The syntactic basis of Masoretic Hebrew punctuation. Language, pp. 28-72.

Aronoff, Mark. 1994. Morphology by itself: Stems and inflectional classes (Linguistic Inquiry Monographs No. 22). Cambridge, Mass.: MIT press.

Arnold, Matthew. 1869. Culture and Anarchy. London: Smith, Elder, and Co.

Baayen, Harald, Piepenbrock, Richard, & Gulikers, Leon. 1995. The CELEX lexical database (release 2). Philadelphia: Linguistic Data Consortium.

Bauer, Laurie, Lieber, Rochelle & Plag, Ingo. 2013. The Oxford Reference Guide to English Morphology. Oxford: Oxford University Press.

Berg, Kristian, Franziska Buchmann, Katharina Dybiec & Nanna Fuhrhop. 2014. Morphological spellings in English. Written Language & Literacy 17.2: 282-307.

Blevins, James. 2013. Word-based morphology from Aristotle to modern WP. In K. Allan (ed.), Oxford Handbook of the History of Linguistics. Oxford; Oxford University Press. 375–396.

[21] Of the many people who have attempted to consciously reform English spelling almost since it was first recorded, the only one to have had any success was Noah Webster. It is worth noting that his greatest success involved a suffix of sorts: he spelled earlier *–our* as *–or* (*flavor* for *flavour*). Aronoff (1978) argues, however, that the original ascendance of *–our* over *-or* was precisely in words where the unstressed syllable was not analyzable as a suffix, making this a very peculiar case of undoing the encoding of a morphological distinction in spelling. Regardless, the reform did succeed.

Bolinger, Dwight. 1946. Visual Morphemes. Language 22 (4), pp. 333-340.

Carney, Edward. 1994. A survey of English spelling. London: Routledge.

Chomsky, Noam & Halle, Morris. 1968. The sound pattern of English. New York: Harper & Row.

Evans, Mel. 2012. A Sociolinguistics of Early Modern Spelling? An account of Queen Elizabeth I's correspondence.
http://www.helsinki.fi/varieng/series/volumes/10/evans/

Flemming, Edward & Stephanie Johnson. 2007. Rosa's roses: Reduced vowels in American English. Journal of the International Phonetic Association 37.01: 83-96.

Fuhrhop, Nanna. 2011. Grammatik verstehen lernen – Grammatik sehen lernen. In: Köpcke, Klaus-Michael & Ziegler, Arne (eds.): Grammatik – Lehren, Lernen, Verstehen. Zugänge zur Grammatik des Gegenwartsdeutschen. Berlin/Boston: de Gruyter, pp. 307-324.

Labov, William. 1994. Principles of Linguistic Change: Internal Factors. Oxford: Blackwell.

Marchand, Hans. 1969. The Categories and Types of Present-day English Word-Formation: A Synchronic-Diachronic Approach, 2nd edition. Munich: Beck

Marchand, Hans. 1969b. Political history and the rise of the suffix /i/ in English. In: Die Neueren Sprachen 15, 353-358.

Martin, Peter. 2008. Samuel Johnson: A Biography. Cambridge, Mass.: Belknap Press of Harvard University Press.

Matthews, Peter H. 1972. Inflectional Morphology. Cambridge: Cambridge University Press.

Saussure, Ferdinand de. (1916) 1959. A Course in General Linguistics (translated by Wade Baskin). New York: Philosophical Library.

Scragg, Donald George. 1974. A history of English spelling. Vol. 3. Manchester: Manchester University Press.

Taft, Marcus. 1979. Lexical access-via an orthographic code: The basic orthographic syllabic structure (BOSS). In: Journal of Verbal Learning and Verbal Behavior 18(1), S. 21-39.

The Helsinki Corpus of English Texts. 1991.

Treiman, Rebecca, and Brett Kessler. 2014. How Children Learn to Write Words. Oxford: Oxford University Press.

Wilcox, Allen R. 1967. Indices of qualitative variation.