Chapter 7

Major lexical categories and graphemic weight

Kristian Berg
Universität Bonn

English spelling has a very interesting regularity: there exists a minimum word length for lexical words. Words of this class have to be at least three letters long, even if they consist of only two phonemes and could be spelled with two letters (e.g., ebb/*eb, egg/*eg). This regularity does not hold for function words (e.g., a, I, be, he, it, etc.). This means that an important distinction between words (lexical vs. functional) is mirrored in the writing system. In a study of Early Modern English texts, I demonstrate that this regularity evolved gradually over the course of 200 years. This is a case of self-organization in spelling: without explicit guidance or regulation, the pattern emerged in usage. The proposed function of having separate constraints for the length of lexical words and function words is a reading aid.

Keywords: English, lexical words, grammatical words, spelling, self-organization

1. How did we get here?

English spelling is morphographic to some extent. This has been demonstrated for a number of phenomena: for example, the alternation between ⟨-our⟩, ⟨-or⟩, and ⟨-er⟩ in British English spelling reflects subtle morphological differences (Aronoff 1978). The form ⟨our⟩ is mostly found with bound bases, and the nouns are mostly inanimate (e.g., colour, favour); ⟨-or⟩ is a deverbal suffix which denotes animate nouns (e.g., mediator, oppressor). This distinction is not mirrored in phonology.

Or take the spelling of English suffixes and word endings (Berg & Aronoff 2017, 2018): we have shown that English spellings became both more uniform (variation in the spelling of word forms was reduced) and more unique (suffixes tend to have distinct spellings that set them apart from homophonic word endings – e.g., nervous vs. service) over time (Berg & Aronoff 2018).

In what follows, I present diachronic evidence for a regularity in the English writing system, the so-called three letter rule. This rule, too, is an instance of

1. I use angled brackets to highlight actual spelling.

https://doi.org/10.1075/cilt.353.07ber
© 2021 The Author
morphographic spelling, albeit a slightly different one. It states that lexical words in English must have a certain minimal length; function words can be shorter. I illustrate the rule and its exceptions (§2), speculate about a potential utilization for readers (§3), and then trace the spelling of two lexical words and two function words through time (§4 and §5). I show that the three letter rule emerged as a result of another spelling change, namely the transition from complex word final patterns like ⟨dde⟩ (as in ⟨madde⟩) to single letter spellings (as in ⟨mad⟩). I then discuss the findings (§6).

2. The three letter rule

The three letter rule concerns the English lexicon, and thus a central part of morphology. It states that English content words must have a minimal length of three letters – even in cases where the regular phoneme-grapheme correspondences would allow for a two letter spelling, as illustrated in (1)

(1) add, ebb, egg, inn, odd

This regularity has often been observed in the past. One of the most prominent mentions is in Jespersen’s Modern English Grammar:

Another orthographic rule was the tendency to avoid too short words. Words of one or two letters were not allowed, except a few constantly recurring (chiefly grammatical) words: a, I, am, an, on, at, it, us, is, or, up, if, of, be, he, me, we, ye, do, go, lo, no, so, to, (wo or woe), by, my. To all other words that would regularly have been written with two letters, a third was added, either a consonant, as in ebb, add, egg, Ann, inn, err – the only instances of final bb, dd, gg, mm and rr in the language, if we except the echoisms burr, purr, and whirr – or else an e […] see, doe, foe, roe, toe, die, lie, tie, vie, rye, (bye, eye), cue, due, rue, sue. (Jespersen 1909: 149)

As Jespersen shows, the three letter rule only holds for lexical words; function words regularly violate it. Lexical words in English have to be of a certain length, and they are “usually bulked up to a minimum of three letters” (Carney 1997: 76). One function may be to avoid homography, as Cook (2014: 70) suggests: “A spoken content word that would logically correspond to two letters is padded out with an extra consonant to avoid confusion with two-letter function words”. Note the similar metaphors used in both citations (bulk up, pad out) – the idea is that words like ebb and odd are made longer than they have to be.2 Historically, that is not what happened (see §5).

2. I would like to thank Rebecca Treiman for this observation.

All rights reserved
A more careful inspection reveals that what is at stake here is not length, but syllable weight, as Evertz (2018) shows. The main argument is that even three letters are sometimes not enough, as in the cases in (2):

\[(2) \ast \text{tre, } \ast \text{sho, } \ast \text{blu, } \ast \text{bla}\]

Apparently, the rime of the graphemic syllable needs to consist of either two vowel letters (\textit{tree, blue}) or of a vowel letter and a consonant letter (\textit{show, blah}). Borrowing established terminology from non-linear phonology, we can say that the rime of monosyllabic lexical words must be associated with at least two (graphemic) moras. Each letter in the rime is associated with one mora. This is the first well-formedness constraint for lexical words, which I will call the ‘rime constraint’. The second one is that lexical words consist of at least three letters (spellings like \langle eb \rangle and \langle eg \rangle comply with the first constraint but are still not heavy enough); I will call this the ‘length constraint’.

In today’s English, there are only a few exceptions to this regularity (Berg 2019: 162ff.). The CELEX database (which contains 52,447 English lemmas, Baayen, Piepenbrock & Gulikers 1995) lists the following two-letter words:

\[(3)\]
\begin{enumerate}
\item[a.] go, ox, ex
\item[b.] mu, pi
\item[c.] do, re, mi, fa, so, la, ti, do, si
\item[d.] ma, pa, mo, po
\end{enumerate}

Among these words, only the ones in (3a) are true exceptions to the graphemic weight minimum for lexical words.3 The words in (3b) and (3c) are of foreign origin (they are Greek letter names and syllable names for the solmization), the ones in (3d) are all shortenings, with \textit{mo} for American English \textit{momentum} or British English \textit{moment}, and \textit{po} for British English \textit{[chamber] pot}, i.e., results of a rather minor word formation type. Not included are \textit{ad} (a shortening), the interjection \textit{hi}, and potentially more counterexamples; however, I assume (a) that the number of such counterexamples is very limited and (b) that they are either the result of minor word formation types or belong to minor lexical classes – i.e., their morphology is special, and so is their spelling.

Thus if we count liberally, there are three exceptions to the minimal weight restriction for lexical words. On the other hand, there are 20 function words among the two-letter words:

---

3. Note, however, that \textit{go} can be used as a function word to mark future tense, and \textit{ex} started out as a Latinate combining form. Also note that the spelling variant \textit{ax} is not included in CELEX. It is remarkable that \textit{ox, ex, and ax} all end with \langle x \rangle; it may be desirable to describe \langle x \rangle as a complex grapheme like \langle y \rangle (see below).
(4)  an, as, at, id, if, in, it, of, on, or, up, us, be, by, he, me, my, no, to, we

With *am* and *is*, there are actually more two letter ‘word forms’ of this type (which do not appear in the results because the CELEX search was for lemmas).

There could be potentially more violations of the weight restriction, however. The list in (5) contains all 57 lemmas in CELEX that consist of two phonemes and that could be spelled with two letters, but which are not:

(5)  add, ass, ebb, eff, egg, ill, inn, odd, off, ore, bee, bow, buy, cue, die, doe, dough, due, dye, fee, foe, ghee, hie, high, hue, key, knee, know, lea, lee, lie, low, lye, mow, nigh, pea, pee, pie, roe, row, rue, rye, sea, see, show, sigh, sow, sue, tea, tee, tie, toe, tow, vie, wee, woe, zee

For some of these words, the two letter spelling exists, but as a homophonous function word (*in*, *be*, *or*, *by*, *no*). All these cases (except for the only function word in (5), *off*) are evidence for the efficacy of the minimal weight restrictions for lexical words in English.

There are more exceptions for three letter words with a single final vowel in CELEX:

(6)  a.  chi, gnu, phi, ski, spa
    b.  cry, dry, fly, fry, ply, pry, shy, sky, sly, spry, spy, sty, try, wry

The words in (6a) are (relatively) recent borrowings; their graphemic structure is special, and so is their status in the lexicon. The words in (6b), on the other hand, are inconspicuous in this regard; they are (for the most part) frequent English words. What sets them apart from other English words is that they end with single final ⟨y⟩. If we want to maintain our claim about the minimal weight of English lexical words, we have to either allow for the exception of ⟨y⟩, or claim that ⟨y⟩ is heavier than the vowel letters ⟨a, e, i, o, u⟩. Either way, the set of exceptions is limited to two very specific types, recent borrowings and words that end with ⟨y⟩.

English content words must fulfill certain weight constraints, while English function words can be graphemically lighter (by the way, the spelling of German words is parallel in this regard). What is the use of this peculiar regularity? In the following section, I will argue that it can potentially be utilized as a reading aid.

3.  Why this regularity?

Minimal weight constraints that only hold for lexical words lead to a set of function words that are ‘lighter’ than lexical words can be. As such, they can be easily identified by the reader. This potentially helps readers with the parsing of sentences: it
tells them which parts are purely grammatical, like the and in, and where to expect lexical content.

To get a better impression of the number of function words that can be identified at first glance, take two short texts from Alice Munro and John Le Carré, and note the words that violate the length and weight restrictions (highlighted).

Alfrida. My father called her Freddie. The two of them were first cousins and lived on adjoining farms and then for a while in the same house. One day they were out in the fields of stubble playing with my father’s dog, whose name was Mack. That day the sun shone, but did not melt the ice in the furrows. They stomped on the ice and enjoyed its crackle underfoot. How could she remember a thing like that? my father said. She made it up, she said. (Alice Munro, Family Furnishings)

Leamas was not a reflective man and not a particularly philosophical one. He knew he was written off – it was a fact of life which he would henceforth live with, as a man must live with cancer or imprisonment. He knew there was no kind of preparation which could have bridged the gap between then and now. He met failure as one day he would probably meet death, with cynical resentment and the courage of a solitary. He’d lasted longer than most; now he was beaten. It is said a dog lives as long as its teeth; metaphorically, Leamas’ teeth had been drawn; and it was Mundt who had drawn them.

(John Le Carré, The Spy Who Came in from the Cold)

A considerable number of running words in texts can be identified as function words on purely graphemic grounds, i.e., without the need for lexical access (29% in Family Furnishings, 27% in The Spy Who Came in from the Cold). We know that readers often skip function words, but rarely lexical words (Dehaene 2009: 17); their length and graphemic makeup can be used as cues to gauge the position of the next gaze fixation (and accordingly, the length of the saccade).

Of course, not all function words can be formally distinguished from lexical words (e.g., and, from, its, about, you, there). However, of the ten most frequent lemmas in the Corpus of Contemporary American English (CoCA, https://www.english-corpora.org/coca/), eight are lighter than lexical words (the, be,4 of, a, in, to (infinitive marker), to (preposition), it).

As with every feature of the writing system that seems to be functional, the question is how it evolved. For the regular correspondences between affix spelling and morphological information, we showed that it evolved gradually in a process of self-organization (Berg & Aronoff 2017). How did the weight minimum for lexical words come about? The remainder of this chapter sets out to answer this question.

4. The forms be, am, and is are lighter than lexical words; the forms are, was, and were are not.
4. Methodology

The data base for this investigation is the large diachronic corpus Early English Books Online (EEBO, https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/). It contains more than 25,000 printed texts between 1500 and 1700, totaling up to 872 million running words. The corpus is neither orthographically normalized nor tagged, and thus it is impossible to search for abstract graphemic patterns (like “consists of a vowel and a double consonant”) or parts of speech. As a proxy, I will instead concentrate on two pairs of words that behave differently in today’s English, *ebb* vs. *web* and *egg* vs. *leg*. As witnessed by *web* and *leg*, stem final /eC/ can be spelled ⟨eC⟩. However, in biphonemic words like *egg* and *ebb*, this is not the actual spelling. The hypothesis is that the spellings ⟨eg⟩ and ⟨eb⟩ are too light; they violate the length constraint.

To obtain the raw data, I use the Kontext interface (https://kontext.korpus.cz) and perform searches for the following different spellings of *ebb*, *web*, *egg*, and *leg*:

(7) a. *eb*, *ebb*, *ebbe*
   b. *web*, *webb*, *webbe*
   c. *eg*, *egg*, *egge*
   d. *leg*, *legg*, *legge*

The results are then annotated for false positives – entries from EEBO with one of the word forms in (7) that are not instances of the lexemes *ebb*, *web*, *egg*, and *leg*, like the following examples:

(8) a. *Wherfore that which is contained vnder the lines AE and EB in commensurable to that which is contained vnder the lines…* (Euclid, 1570)
   b. *The time is whilſt Thomas Webbe is Preaching in the publique Congregation…* (Stokes, Edward, Esq., 1653)
   c. *The Subftilar line EG being aw the line XW through the center C…* (Aspley, John, 1662)
   d. *Colonell Robert Legge, Governour of the Town, …* (Vicars, John, 1646)

After filtering these entries, there were 30,454 word forms left, which are analyzed in the next section.
5. Results

The 30,454 forms are distributed as follows among the different lexemes and spelling forms (see Table 1):

Table 1. Absolute (and relative) frequencies of the words *ebb*, *web*, *egg*, and *leg* in three different spellings in the EEBO corpus

<table>
<thead>
<tr>
<th></th>
<th>(C)</th>
<th>(CC)</th>
<th>(CCe)</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ebb</em></td>
<td>122 (3%)</td>
<td>1558 (41%)</td>
<td>2113 (56%)</td>
<td>3793</td>
</tr>
<tr>
<td><em>web</em></td>
<td>2519 (72%)</td>
<td>192 (6%)</td>
<td>763 (22%)</td>
<td>3474</td>
</tr>
<tr>
<td><em>egg</em></td>
<td>249 (2%)</td>
<td>4435 (44%)</td>
<td>5428 (54%)</td>
<td>10112</td>
</tr>
<tr>
<td><em>leg</em></td>
<td>8106 (62%)</td>
<td>1280 (10%)</td>
<td>3689 (28%)</td>
<td>13075</td>
</tr>
</tbody>
</table>

The total frequencies of *ebb* are in the same order of magnitude as those of *web*, and the same holds for the *egg/leg* pair. As for the form variants, the simple coda spelling is only a minor variant for *ebb* and *egg*, while it is the dominant variant for *web* and *leg*.

The more interesting question, of course, concerns the development over time. Figure 1 shows the ratio of today’s form among all spellings for the four lexemes (there are no data for *egg* and *web* in 1510, hence the white space).

![Figure 1](image-url)

Figure 1. Relative number of word final spelling patterns ⟨CCe⟩, ⟨CC⟩, and ⟨C⟩ among all *leg/ebb/web/egg*-spellings in the EEBO corpus
For *leg* and *web*, we see a gradual transition from ⟨legge⟩/⟨webbe⟩ to ⟨leg⟩/⟨web⟩ between 1500 and 1700. The relatively long duration of this transition makes it improbable that it was governed by spelling authorities; it is more likely an instance of self-organization. The spellings with double final consonants ⟨legge⟩/⟨webbe⟩ appear in the 17th century, but they remain marginal for both lexemes and never really catch on. Instead, as Aronoff (p.c.) observed, this pattern is regularly used in the spelling of surnames that would otherwise be homographic with content words: *Webb, Kidd, Scott, Crabb, Robb, Trapp.*

Interestingly, there is a considerable amount of variation within texts. The variation we see on a global scale (e.g., in Figure 1) comes about (at least in part) because writers use both forms in their texts, as the examples from *A Relation Of Two several voyages Made into the east-indies* (by Christoph Frick, written 1700) show:

(9) [...] for when the others see one of their company thus caught by the Legg, they won’t come near the Coco-Nuts all that day (p. 45)

(10) [...] I was immediately shot in the Leg with one Arrow, and with another in my Thigh (p. 312)

Between 1550 and 1640, in 59% of the texts in which the spelling of *leg* may vary (i.e., texts which contain more than one form of the lexeme), it does actually vary: 59% of these texts are inconsistent with regard to the spelling of *leg*. For *web*, the ratio is 53%. That means that the transition from one form ⟨legge⟩ to another ⟨leg⟩ did not proceed along the lines of texts (with each author consistently using their favorite spelling), but that it involved variation within texts as well. This is a situation we expect to see in a self-organizing system.

For *ebb* and *egg*, the transition is much quicker, and it only starts when *leg/web* have mostly switched to their modern forms (around 1640). Here we go from ⟨egge⟩/⟨ebbe⟩ to ⟨egg⟩/⟨ebb⟩. The simple coda spelling ⟨eg⟩/⟨eb⟩ was never a serious contender. There was a short period of time where *leg* and *web* were already mostly spelled ⟨leg⟩, ⟨web⟩, but *ebb* and *egg* were still spelled ⟨egge⟩ and ⟨ebbe⟩.

Why did these spellings have to change at all? After all, ⟨egge⟩ and ⟨ebbe⟩ fulfill both minimal word constraints. My hunch is that double consonant + ⟨e⟩ spellings (as in ⟨legge⟩ and ⟨webbe⟩) were soon regarded as markers for the old way of spelling, the cumbersome, baroque way, which was replaced by the fresh new phonographic simplicity of ⟨leg⟩ and ⟨web⟩. Similarly, the variants ⟨egge⟩/⟨ebbe⟩, I suppose, looked out of fashion and thus had to change. And apparently, ⟨eg⟩/⟨eb⟩ (analogous to ⟨leg⟩/⟨web⟩) was not an option because it left too little graphemic substance.

Both pairs *egg/ebb* and *leg/web* correlate closely with respect to the ratio of new forms over time (Pearson’s $r$ (ebb/egg): 0.98, $P < 0.001$; Pearson’s $r$ (web/leg): 0.88, $P < 0.001$). That means even though I only investigated two lexemes per category,
these two words show a high degree of similarity when it comes to the transition from one form to another, and it is likely that a third lexeme of the same pattern will behave in a similar way.

6. Discussion

In today’s English, *leg* and *web* show a pattern that is distinct from phonologically similar words like *egg* and *ebb*: the spellings ⟨leg⟩ and ⟨web⟩ are phonographic spellings, while ⟨egg⟩ and ⟨ebb⟩ are not. The double consonants are usually explained with reference to a graphemic minimum that English content words must fulfill. As noted in §2, this is often phrased in terms of ‘padding’ or ‘bulking up’. But historically, we find the opposite development: *egg* and *ebb* were dominantly spelled ⟨egge⟩ and ⟨ebbe⟩ until around 1650. After that, they dropped their final ⟨e⟩. These spellings were thus abridged, not padded – it is only because other words were abridged even more that we see the difference between the groups of lexemes today. We can thus pin down the origin of minimal word constraints in English to the second half of the seventeenth century.

Minimal word constraints are of course not exclusive to written English; there are analogous phonological (prosodic) constraints as well. In spoken English, lexical words are minimally prosodic words, which consist of at least one foot with at least one syllable with at least two moras (see, e.g., McCarthy & Prince 1994, Selkirk 1996). Function words, on the other hand, are often realized as prosodic clitics (Selkirk 1996). But the reflexes of this lexical difference are not isomorphous: in spoken English, function words like *at* regularly appear in reduced/weak forms – but in phrase-final position, they are never reduced (*What did you look at yesterday?*, Selkirk 1996: 200). In written English, the difference is categorical: there is no formal variation. Thus even though both written and spoken language are sensitive to the distinction between function words and lexical words, we cannot derive one from the other because the actual implementation differs.

The same holds for the potential use of the respective minimal word constraints. I suggested above that minimal words in writing could help provide the reader with cues to lexical access, and to potential points of gaze fixation. Selkirk (1996: 187) hypothesizes that the prosodic differences between function words and lexical words could be “exploited by the language learner” whose “first order of business is which words are functional and which are lexical” – a very different function for the same contrast between lexical words and function words we find in speech and writing. This is not surprising: the same grammatical difference has different reflexes in the spoken and written variety, and with them vary the
potential uses for the reader/hearer. A simple derivation of written from spoken language cannot do justice to this complex relationship. The English writing system, just as its phonology, is sensitive to this lexical distinction, and spelling thus involves linguistic analysis.

Acknowledgements

I would like to thank the editors of this volume for valuable comments on an earlier version of this chapter, and I would like to thank Mark Aronoff for discussing the three letter regularity regularly with me. On a personal note, Mark is one of the kindest and most generous colleagues I know, fueled by the desire to understand how language works, and never dogmatic. He is truly Jacobsonian in the scope of his interests – nothing linguistic is alien to him. Speaking of Jacobson, I have always wondered whether Mark’s interest in writing was (at least in part) in some way attributable to the Prague circle (with his mentor Morris Halle as the proxy). Thank you, Mark!

References