English consonants are regular (except C and G)

Anne Vainikka/The Verb Company and Johns Hopkins University

January 2016, draft

An analysis of the English spelling system for the 20,000 most common words of English reveals that the majority of the English graphemes are, unexpectedly, nearly 100% regular, similar to the Finnish graphemes.

The table on the next page presents all the English consonant graphemes except those involving the letters C or G. Each grapheme is counted twice: once word-initially, and once elsewhere in the word; some items have been combined for space reasons. The first data column shows whether the consonant (in that position) allows for the option of omitting the consonant, that is whether the consonant can be silent (e.g. for B ‘other’, the word DEBT shows that such a B can be silent).

The second data column in the table shows whether the grapheme has more than one variant (excluding the silent variant), and the third column shows whether the grapheme is always read the same way; if not, the exceptional words are listed (from the set of 20,000 most common words). Words that would have been exceptional but were excluded from the table because they fell below the cut-off point of 400 occurrences in COCA (occurrence in COCA in parentheses) were: *mnemonics (63), quay (267), brusque (312), pique (353), chintz (227), quartzite (82), azure (394), brazier (161) and glazier (94).* Excluding these words from the analysis allowed for a simpler system. Otherwise all words were covered from various reference sources, with a final check of the very comprehensive reference book of Bishop (1986).

	empty option	more than one variant	exceptions
B: initial + other	YES	NO	NO
D: initial	NO	NO	NO
D: other	NO	YES [d t j]	NO
F: initial	NO	NO	NO
F: other	NO	NO	‘of'
H: initial + other	YES	NO	NO
PH/RH/SH initial/other	NO	NO	NO
TH: initial	NO	YES	‘thyme’
TH: other	NO	YES	NO
WH: initial	YES [W silent]	NO	NO
J - initial + other	NO	NO	NO
K: initial + other	YES	NO	NO
L: initial + other	YES	NO	‘colonel'
M: initial + other	NO	NO	‘comptroller'
N: initial + other	YES	NO	NO
NK: non-initial	NO	NO	NO
P: initial + other	YES	NO	NO
PH: initial + other	NO	NO	NO
QU: initial	NO	NO	NO
QU: other	NO	YES [kw k]	NO
R: initial + other	NO	NO	NO
S: initial	NO	NO	NO
S: other	YES	YES [s z sh zh]	NO
T: initial	NO	NO	NO
T: other	YES	YES [t sh ch]	NO
V: initial + other	NO	NO	NO
W: initial + other	YES	NO	NO
X: initial	NO	NO [z]	NO
X: other/Z: initial	NO	NO	NO
Z: other	NO	NO	pizza;waltz,quartz, rendezvous

Given the table, we will now attempt to quantify the regularity of the English consonants based on the table, excluding C and G. The table, in fact, covers both short and longer words of English (among the 20,000 word set). Counting the initial vs. non- initial graphemes as separate graphemes, we end up with the 48 graphemes shown in the table. Excluding the silent version for the moment, six of them have more than one version (TH initially and elsewhere, and D S T QU non-initially). 6 out 48 equals 12.5%; thus, 87.5% of these graphemes are regular and fully transparent, as they are always read the same way, as long as we exclude silent consonants.

If we now take the approach that reading should begin with short words, and if we can further exclude the inflectional affixes (-s, -ed, -ing), the picture in the table becomes even more regular. When we consider only (underived) short words, we get the following situation: **D and T are regular everywhere; TH has the voiced and voiceless versions everywhere; QU is regular everywhere, but there are 3 new exceptional words: clique, mosque, torque; non-initial S still has a voiced and voiceless version.**

This means that only three graphemes are 'exceptional': (1) the initial TH, (2) the non-initial TH, and (3) the non-initial S. That is, 45 out of 48, or 94%, of the consonant graphemes in short words (without a suffix) are regular; in addition, much of the distribution of S and TH is also predictable.

What remains are the exceptional words in the last column. When considering only short words, we end up with this set of seven exceptional words: ***of, thyme, waltz, quartz, clique, mosque, and torque***. Among the 20,000 most frequent words of American English, these are the only short words in which a consonant (other than S, TH, C or G) is pronounced in an unexpected way. In addition, we need to include the silent consonants. As shown

on The Verb Company website (Word Wheels 2, 3a, and 3b), there are 136 short words (in the 20,000 set) of English that contain a silent consonant. That is, adding up the two sets of exceptional words, we end up with 143 (136+7) such words. While the analysis deals with the 20,000 most frequent words of English, only an estimated 4,800 of them are short (based on COCA). The proportion of exceptional short words, then, is 143/4800, or 3%.

Summarizing, in the 20,000 word set, and excluding C and G, 94% of the consonant graphemes are completely transparent and predictable (S and TH being the exceptions). Given the 45 regular graphemes, 97% of the short words are regular, completely predictable and transparent; even the 3% of exceptions mostly involve a silent consonant. Thus, the set of English consonants without C and G approaches the transparency and regularity of the Finnish spelling system.