Regularity in the phonemic subsystem of English spelling: A study of the monosyllables

Brett Kessler*

December 25, 1995

Abstract

A list of 4,724 English monosyllables was studied in an effort to determine the degree of regularity in writing words according to the standard orthography of American English. The spelling and pronunciation of the words were aligned semi-automatically, and these alignments were used to gather data on how individual phonemes are spelt in various phonetic environments. The data were also used to build a set of rules for reducing English words to writing; the complexity of these rules was then used as another measure of regularity.

1 Background

The irregularity of English orthography, when viewed as a system for mapping phonemes to letters, has been noted for centuries. Almost all commentators on English spelling have noted this irregularity and blamed it on the conservatism of an orthography which has not kept pace with the phonetic evolution of the language. The failure of spelling reformers to convince the public of the need for reform has in general been explained by non-linguistic factors such as the unwillingness of already literate adults to inconvenience themselves for the sake of their struggling children, the expense of retooling typewriters and printing presses, and the feared inaccessibility of pre-reform writings to those who have mastered only the revised spelling rules. Perhaps the ultimate estimation in this tradition is that of [Sampson 1985], who claims that present English orthography might best be characterized as fundamentally not phonological but logographic, assigning essentially arbitrary spellings to each morpheme. On the other hand, with the advent of generative linguistics, many prominent researchers in recent decades have taken a contrarian stance, holding that English spelling is linguistically quite regular. [Chomsky 1970], for example, approvingly

*The author gratefully acknowledges the consultation of Rebecca Treiman of Wayne State University, to whom the idea for many of these measures and analyses is due.

quotes [Chomsky 1970?], who held that it "corresponds closely to a level of representation that seems to be optimal for the sound system of a fairly rich version of ... spoken English," the purported irregularities becoming regular when one considers deeper levels of the phonology and the syntax. [Venezky 1970] similarly held that the irregularity in English spelling, if properly considered as a morphophonemic system with sufficiently complex phonotactic rules, is mostly a matter of unnaturalized loan words.

I was interested in the degree to which the regularity of English phoneme to grapheme correspondences could be quantified. Although rather comprehensive studies of the writing system have been published, including the cited work by Venezky, hard figures are rare, even though Venezky based his work on a machine-readable list of twenty thousand high frequency words. A noteworthy exception is [Hanna 1966], although that book gives only type, not token counts, and is restricted to moderately common words. For the present purposes, it seemed particularly advantageous to find a way to factor out the morphophonemic factors that are clearly present in English spelling (for example, the different pronunciations of the *a* in *sane* and *sanity*). Since most of the orthographically levelled variation is found in polysyllabic derived forms, I worked with a comprehensive list of the monosyllabic words current in standard American English, excluding inflections and derivations (notably *th* forms) where the stem does not change. This tack also had the advantage in resulting in a smaller (4,724 word) corpus, providing a reasonable chance for decent quality control in a short time span.

2 Word list

The word list consists of all monosyllabic words found in the most recent unabridged Random House dictionary [Random House 1987], provided that there was no indication that the word was archaic or obsolete, slang, nonstandard, regional, paralinguistic (interjections like *sh!*), foreign, or not in general use in the United States of America. Regular inflected or derived forms (*calls, called, warmth)* were omitted, although irregular forms were included (*does, did, health),* counting as "irregular" anything not based on concatenation to the full free form of the base word.[1] Initialisms and words regularly spelt with any punctuation or accent marks were excluded. Proper names were included only if they were significantly Anglicised (*Spain),* or common given names (*Jane).* This led to the perhaps surprising result that names of American places and famous people were omitted, whereas foreign place names like *Rhodes* were included in this list of American words. The reason for this is to some extent the

[1] These somewhat irregular forms were included because throwing them out would have resulted in omitting many stems that are clearly spelt phonemically; it is apparent, for example, that the spelling *taught* is not unduly influenced by the spelling of *teach,* and the same seems to hold true for most irregular past forms. On the other hand, this policy resulted in including a certain number of instances of morphophonemically-spelt morphemes, the only significant example being those spelt *-s.* A replication of this study would do well to delimit those derivational and inflectional morphemes so that they can be excluded from the tallies.

uncertain status of the naturalization of American names, so many of which are of non-English origin, and also the encyclopedic nature of the source dictionary, which includes many proper names not likely to be familiar to most writers. Nevertheless this policy is to some degree arbitrary, and a replication of the study omitting proper names might be useful.

2.1 Details

As a convenience for me as human editor, a base word list was formed by combining words found in several other word lists, including a list supplied by Venezky, and the list published by [SCRL 1973]. However I corrected this list by manually adding or deleting words based on the authority of the Random House dictionary, whose decisions on inclusions, spelling, pronunciation, and usage were considered absolute.

All monosyllabic words in Random House were included except for the following:

– Those having a regional indicator showing chief region of usage being outside the US: e.g., *Scot., Brit.,* even if the term is qualified by *Chiefly.* Example: *rax.*

– Words qualified as being in a different language: *ach, German.*

– Certain classes of words that almost always retain foreign spellings: terms from non-English monetary systems (coins, monetary units), weights and measures (liang), calendars (names of months), letters of the alphabet, peoples and nations. I include however such terms if they are manifestly Anglicised, such as by having a different spelling than in the foreign language. Terms from Latin or Greek antiquity are automatically considered Anglicised.

– Words printed in italics.

– Initialisms spelt with all caps: A, AID. I include only initialisms that now look and are pronounced like normal words *(Wac).*

– Words spelt with punctuation marks, including periods, hyphens, and apostrophes.

– Words with accent marks.

– Prefixes, suffixes: a-, -a

– Words found as parts of multi-word entries: ad lib

– Secondary spellings: adze, aesc, Ab. That is, I exclude words that are defined exclusively by reference to another spelling, and instead include that spelling. However, if a word appears in two different spellings and each of those spellings has a different main pronunciation, I consider them to be two different entries and include both, even if it is obvious that the words have a common origin and meaning.

– Words with non-English sounds in their pronunciation: Ain [aN]. These symbols are spelt with caps: A, KH, N, OE, R, Y.

– Words not containing any indication of the time when they entered the language. (These appear universally to be only marginally English.) Exceptionally, proper given names are not discarded under this rule.

– Words marked *Nonstandard, Slang, Obs[olete], Archaic, Rare, Older Use, Facetious, Baby Talk, Literary, Eye Dialect, Pron. Spelling.* However, I do include terms with other labels, such as those indicating usage within a particular subject field (e.g., *Chemistry),* and such labels as *Informal, Vulgar, Disparaging* and *Offensive* (if not also *Slang,* etc.)

– Words whose definition indicates that they are used only in non-English contexts or names, e.g. *de.*

– Compounds of productive morphemes: *acts, paid.* In particular, if the pronunciation of a plural or verbal 3rd person singular is formed by adding /s/ or /z/, I omit it; if a past tense or participle is formed by adding /d/ or /t/, I omit it; if a noun is formed by adding /q/ (e.g., *warmth),* I omit it. This holds even if the spelling seems irregular *(pay, paid).* However, if the pronunciation of the base changes in the process, I include the word *(catch, caught; have, has; long, length).* Also, if a word looks like a plural or past participle but there is no entry for the base word, I include the word.

– Most proper names, such as people's surnames, place names. However I do include given names that have their own entries as such, e.g., *Bob,* a male given name.

Words that have the same spelling and pronunciation are considered to be one entry for the list, even if there are different meanings, or even if there are multiple entries in the dictionary. (For example, the dictionary has 3 entries for *bill* in American usage, but they are all pronounced the same, so I give it one entry.) However, if they differ in capitalisation, I include both (e.g., I also include the given name *Bill* as a separate entry), even if the common and proper words seem to have the same etymology.

On the other hand, a single Dictionary entry may give different pronunciations for different word senses. In this case, I make two different list entries.

If a dictionary has multiple spellings or pronunciations for a word, the filtering criteria are applied to the first such. For example, if the first pronunciation is polysyllabic or includes a foreign sound, the word is excluded, even if a subsequent pronunciation is monosyllabic or nativized.

Words whose pronunciations include a stress mark are omitted as polysyllabic, even if there is only one vowel: *button* [but'n].

2.2 Word list entries

The word list includes the spelling, pronunciation, and frequency of each word. (Its age is also included, but this information was not made use of in this study.)

The first spelling of the main entry in Random House was used; other spellings were included only if a different pronunciation was associated with it, and then the pair was listed as a separate entry. The spelling included initial capitalization if appropriate, although case was not considered distinctive in the actual analyses carried out. By the rules of exclusion, no spelling had any characters other than the twenty-six letters of the alphabet, all letters but the initial being in lower case.

The pronunciation was taken from the first listed in the dictionary. The transcription was based on a common scheme for approximating the International Phonetic Alphabet in the impoverished ASCII code space. Lower case symbols are close to the expected IPA value, whereas majuscules represent arbitrary divergence from the corresponding minuscule. The following table describes how the dictionary symbols were transcribed on the word list. The sample words will serve to define the phonetic values of the symbols in my transcription system.

RHD	Examples	ASCII
a	act, bat, marry	a
ā	age, paid, say	e
â	air, dare, Mary	E (like e, below)
ä	ah, part, balm	A
b	back, cabin, cab	b
ch	child, teacher	T
d	do, madder, bed	d
e	edge, set, merry	E
ē	equal, seat, bee, ear, mere	i
f	fit, differ	f
g	give, trigger	g
h	hit, behave	h
hw	which, nowhere	H
i	if, big, mirror	y
ī	ice, bite	Y
j	just, tragic	D
k	keep, token	k
l	low, mellow, all	l
m	my, summer, him	m
n	now, sinner, on	n
ng	sing, Washington	G
o	ox, bomb, wasp	O
ō	over, boat, no	o
ô	order, ball, raw	c
oi	oil, joint, joy	C
oo short	book, tour	U
oo long	ooze, fool, too	u
ou	out, loud, cow	W
p	pot, supper	p
r	read, hurry	r
s	see, passing	s
sh	shoe, fashion	S
t	ten, matter	t
th	thin, ether	q
th slash	that, either	Q
u	up, sun	V
ûr	urge, burn, cur	R (eg: bûrn = bRn)
v	voice, river	v
w	witch, away	w
y	yes, onion	j
z	zoo, lazy	z
zh	treasure, mirage	Z

As can be seen, only quite minor deviations were made from the pattern employed in the Dictionary; indeed, almost all changes were the direct substitution of one symbol for another. The main exceptions were the omission of the tiny schwa used in diphthong + /r/ sequences, which is clearly subphonemic; *ûr,* which was probably meant to be interpreted as two symbols but which I translated to one, /R/ (since *û* only appears before *r* and the sound is phonetically unitary); and *â*, which I treated as a phonetically conditioned variant of *e* (since it only occurs before *r,*[2] where the identical or similar sound *E* never appears).

The pronunciation transcribed is a cross-dialectal generalization, although there are probably Americans whose phonemic inventory could be accurately described by it. Some important qualities of the "dialect" described in the word list are listed below. Most of these variations in accent are quite regular, and it may be worthwhile replicating this study with adjustments more closely approximating more widespread modern dialects.

– There are two low back vowels, /A/, and /O/, as in *spa* and *cot,* although /A/ is fairly marginal (except before /r/, q.v. infra). For most Americans, these phonemes fall together into one phoneme, except before /r/, where another merger takes place. Thus they have only one low back vowel. This is probably the single greatest caveat to keep in mind when considering the linguistic or psychological verisimilitude of this transcription for any one speaker.

– Similarly there is a distinction between /O/ and /A/ on the one hand and /c/ on the other *(cot, caught),* which is rapidly being lost in America, though this has not yet progressed to the same extent as the previous merger. So for many speakers, I am transcribing their single vowel as three.

– /r/ is sounded where the letter *r* appears, and there is no intrusive (linking) /r/. The distinction /o/ vs. /c/ vs. /O/ is neutralized before /r/, appearing as /c/ (*horse* = *hoarse,* /hcrs/). Similarly historic /er/ is lowered to /Er/ *(care* = /kEr/) and /ur/ to /Ur/ *(boor),* but /ir/ is so transcribed (not /yr/), as in *beer.* In contrast with Scottish English, the early Modern English sequences /Er/, /yr/ and /Ur/ (viz. /Vr/) are all neutralized to one vowel merging with the /r/, giving /R/.

– The word list recognizes a voicing contrast in /w/ vs. /H/, *witch, which,* which is also being rapidly lost in the recent couple of generations.

– Where there is some variation between /a/ and /A/ in such words as *half, can't* and *aunt*, this list gives /a/.

– /ju/ is transcribed for the sound that many speakers pronounce as centralized /yu/ as in *feud.* Note that it is treated as a sequence of two phones. The /j/ of /ju/ is dropped after coronals, as in *new* /nu/ and *due.*

Finally, in order to give some idea of how common the words are in practice, the token count from [Francis 1982] is added. This is a frequency analysis of a corpus of approximately one million words of text published in 1961. Unlike most computerized word counts, this has been disambiguated by word class, which usually suffices to distinguish differently pronounced homographs. In the very few instances where such homographs are of the same word class and there-

[2] It also occurs finally, in the word *yeah,* but this too is distinctive, since /E/ never appears there.

fore have a merged frequency in [Francis 1982], the frequency was split between the homographs, evenly if I had no intuition that one word was overwhelmingly more frequent. In the word list, I treated each word as a particular inflected form, not as a lemma; so the frequency count for *cat* gives the frequency only of that base form, excluding the frequency of *cats, cat's, cats'.*

3 Alignment

The problem of aligning spelling and pronunciation, that is, indicating explicitly which letter or letters spell which sounds, is not as straightforward as it would appear at first consideration. Everyone would agree that in the word *kit,* the three letters spell the three sounds /kyt/, in order. But problems arise in more complex words, in the end raising questions even concerning this uncontroversial explication.

– "Silent *e.*" In words like *kite,* one traditional explication is that the final *e* is silent, but makes the vowel long. Another is that *i* and *e* together constitute the /Y/ sound. Both of these assume therefore that the order of graphemes in spelling can vary from the order of the phones. A more historical account would say that the *e* is a relic of the fact that the syllable was previously open, and that that openness accounts for the tense vowel; therefore it is a mark that the *t* behaves as if it is in the onset of a following syllable rather than the coda of the previous. Or to simplify somewhat, *te* spells /t/ and also conditions a tense pronunciation of the preceding vowel. These accounts are equivalent and of little direct interest to the graphonomist, but make an enormous difference when it comes time to explicitly align spelling and pronunciation.

– Conditioning. In words like *cite*, the *c* spells /s/ and the *i* /Y/, but this does not happen in a vacuum. The *c* spelling is licensed by the fact that the succeeding letter is an *i;* /s/ may not have that spelling except before *e, i,* or *y.* In a very real sense, then, the *i* is part of the spelling of the /s/.

– Silent letters versus polygraphs. Certain spellings such as *sh* for /S/ are commonly explicated as digraphs: both letters together spell the /S/ sound. In words like *debt* on the other hand, the *b* is consistently explicated as a silent letter, and not part of an *eb* or *bt* digraph. While these particular intuitions seem clear, multitudinous boundary cases occur. In *night,* the *gh* may be part of an *igh* trigraph, or it may be a silent digraph that nevertheless conditions the *i.* In *bought,* it is quite unclear which letters are part of a polygraph and which are silent. If *o* is /c/ as in *soft,* then perhaps *au* is not a digraph in *sauce.* If *ou* spells /c/, then it is curious that this never happens except before *gh.* If *ough* spells /c/, then perhaps *gh* should always be considered part of polygraphs, even in words like *caught* where **caut* would be a regular spelling. There do not seem to be any firm criteria for distinguishing the two cases. And indeed, there does not seem to be any firm reason for not saying that in *ship,* the *h* is silent but conditions the pronunciation of the *s.*

– Etymology. Those familiar with a word's etymology are unlikely to tolerate sound/letter assignments that flout the historical state of affairs. In particular, most would agree that recently borrowed words that retain their original spellings should have assignments that correspond, within the limits of the nativized pronunciation, to the presumed sound/letter alignments of the original language.

Confronted with these issues, I adopted the following policies:

– Spelling is strictly sequential; alignment lines do not cross. This settles the issue of whether silent letters might not be put to good use helping spell discontiguous phones in the word. In particular, final *e* after consonants in syllables with tense vowels cannot be considered as part of the spelling of the vowel.

– No many-many alignments. It may be that one phone is spelt by multiple letters (*she* = /Si/) or that multiple phones are spelt by one letter (*six* = /syks/), but multiple phones cannot be spelt by multiple letters. More strictly, a phone may be associated with any number of contiguous letters, but only if none of those letters spells another phone in whole or in part. Thus in cases of conditioning, a choice must be made as to which letter spells what; we cannot say that in *cite* *c* and *i* both spell /s/ if *i* also spells /Y/. Conditioning is considered a secondary factor that can be explored once the primary associations have been made.

– No unassigned letters. All letters must be assigned to one or more phones; none are analyzed as "silent." Therefore in analyzing words like *bought,* the *gh* question reduces to one of whether those letters should be considered part of the vowel or part of the following *t,* a question on which there are clearer intuitions. Of course this decision does not literally mean that there are no silent letters; rather, that all letters in some way contribute to the phonological spelling. In particular, final silent *e* after a consonant is considered part of the spelling of the consonant. In analyses later on, we can see to what extent *e*-less and *e*-ful spellings of consonants condition the spelling of the vowel.

Although it would be nice to write a programme that could align spellings and pronunciations automatically, given no data beyond the word lists and these general policies, I abandoned the idea after a few tentative attempts. Etymology is one confounding issue; although a human brings to the task full cognizance of the spelling rules for Latin, French, and other languages from which words have been borrowed wholesale, to the computer these are often perturbations whose rarity causes them to become lost in the forest. Another factor is that the spelling system as a whole is a good deal more textured than might be immediately apparent; it certainly cannot be described by frequency models. For example, one programme insisted that in *Zeus, Ze* spells /z/ as in *blaze, doze,* etc., and *u* spells /u/ as in *rude, lube,* etc. Certainly those spellings are more frequent than any *eu* = /u/ correspondence in English. And yet any human would prefer the latter. This is due partly to etymology, partly to greater sensitivity to context, and partly to a higher understanding as to the functional utility of the various spellings. I do not think the latter two are beyond the scope of a computer programme, but it is not immediately clear to me how to implement them, at least not in a short time frame.

Instead I took a semi-automatic approach. I prepared a list of 292 grapheme to phoneme correspondences operative in the monosyllables list. See for example the segment reproduced below[3]. The third column, an example, is a comment for the human reader. A programme uses these rules to align the words from left to right, so the ordering of the rules is significant. For example, the programme will attempt to match *aye* before attempting to match plain *a,* but both are attempted before *ach.* This occasionally resolves some ambiguity in whether a letter should be assigned as part of a polygraph to its left or its right.

```
aye Y aye
aye E crayer
aye e raye
a e a
a a abb
a A al
a c all
a E are
a O quad
a V was
ach a drachm
ach O yacht
```

Rather surprisingly, in view of this simple control structure, the programme produced a satisfactory alignment of all words, without the need for human post-editing. Of course, this by no means is to say that there are not any alignments that different people might disagree on. The following snapshot shows part of the monosyllables list with alignments added.[4] The first three columns are the pronunciation, spelling, and frequency, as previously discussed. The last two columns show the alignment, by padding with dashes when there is a one-many correspondence. Again, it is perhaps surprising that this technique gives a satisfactory alignment, since there is no indication as to whether the character should form a cluster to the left or to the right. The convention used here is that it always clusters to the left (except of course at the beginning of the word). That this works perhaps indicates that English orthography, at least in the purely phonemic module, evolved to minimise or even eliminate parsing ambiguities, despite the extensive use of digraphs.

```
blRt blurt 0 blurt blR-t
blVS blush 2 blush blVS-
blYQ Blythe 0 blythe blYQ--
bcr boar 1 boar bc-r
```

[3]The entire list is given in the Appendix as file `rules`.

[4]The entire file, called `dictAlign`, can be got via anonymous ftp from hplnls.hpl.hpl.COM under dist/bkessler/spellings/data/

```
bcrd board 239 board bc-rd
bost boast 8 boast bo-st
bot boat 73 boat bo-t
bOb bob 0 bob bOb
```

This file was used as the data file for all subsequent analyses. Lisp programmes read in the entries and generate various reports on what letters spell which phones in various environments.

4 Analyses

The basic method of analysis performed is the `plain-report,` which is included in full in the Appendix. Here is the automatically-generated analysis of the phone /i/:

```
      TYPES  TOKENS

  /i/
ea     148    5785 Bea beach bead beak beam bean beard beast beat bleach
ee     144    5492 bee beech beef beep beer beet beeves bleed bleep
e       42   95801 be cede cere cete Cleve deke deme eke eve Eve gene
ie      35     721 bier Brie brief chief fief field fiend fierce frieze
i       21      45 chic clique fine firn gi gigue guige kirsch li lisse
ei       8      46 ceil Keith Neil Reid seize sheik weir weird
ay       1       0 quay
es       1       0 mesne
eigh     1       3 Leigh
eh       1       0 lehr
ey       1      85 key
      ----  ------
TOT    403  107978
H    2.149   0.664
```

All the different spellings of the /i/ phoneme are listed in the leftmost column. The next column tells how many different words they appeared in. The heading `TYPES` refers to word types, without reference to the number of times they appear in text. The third column tells how many times the correspondence appears in running text, in the million-word Brown corpus. The line is filled out with as many examples as will fit. At the bottom these per-type and per-token counts are totaled, and the uncertainty statistic H is given. This is a probabilistic measure of how easy it is to guess a spelling; the more spellings possible and the less distinguished they are by probability, the higher the H measure. Note here that in the tokens column, where the spelling *e* becomes wildly more probable than the other spellings (because of the frequency of words like *be,* *she, he,* and *me*), the H measure drops significantly vis-à-vis the types column, which is much smoother and whose top two entries are virtually indistinguishable probabilistically.

This might be a good opportunity to mention the differential utility of the types and token counts. It is of course of great importance to know the general frequency with which words appear in actual text. A spelling that appears in five word types that appear several hundred times in the Brown corpus is in some sense more important than one that appears in five hapax legomena, and an educator may get more mileage out of teaching the totally anomalous spelling *ey* as in *key* before teaching *i,* which appears in twenty-one different words. In general, however, I am concerned with the overall spelling system, as it appears to the fully literate adult who knows all the words in the corpus. One can assume that when the adult comes to spell a word, s/he brings to the task the assumption that a given word is always spelt in the same way, and the capability of associating certain exceptions with the appropriate word. In this context, the token count is irrelevant.[5] In a rule-based system there is no utility in giving prominence to the /i/ - *e* correspondence, even though it works for nine out of ten word tokens one tries to write, because one needs rules for coping with words that are not so spelt. And by the time one finishes the list of subsidiary rules, which are based on such matters as type of environment, and exceptions, which must be itemized by word type, one finds there is no longer any advantage to giving any pride of place to the *e* rule.

It turns out that /i/ is perhaps the most difficult of all phones to spell. The average values are given below:

[5] One might observe that adults in general do not know all the words in the word-list, and that the frequency might profitably be used as a filter for relegating some words to a marginal status. Unfortunately, a one-million word corpus is not large enough for this task. Such common words as *ewe* and *brooch* are not found at all in the Brown corpus.

```
Phone types: 41
                       WORD TYPES     TOKENS
Phone tokens:             16409      1543854
Ave phone tokens/type:     400         37655

Ave H:                    1.167      0.789
Ave (weighted) H:         1.095      0.681
```

The average H just sums the H values of each phone and divides by 41, the number of phones. The weighted average weights each H by how frequent the phone is, and so probably gives a more accurate picture of the general difficulty of spelling. Of course here the difference is quite tiny.

Taking advantage of the convenience of the uncertainty statistic H, I computed the uncertainty of spelling phones occurring in various places in the syllable. In the table below I present only the weighted H; full reports on each phoneme are given in the Appendix for each test.

	WORD TYPES	TOKENS
Phonemes	1.167	0.789
Initial phones /#	0.373	0.154
Final phones /_#	1.301	0.654
VOWELS	1.392	0.963
Initial vowels /#_	1.113	0.159
Medial vowels /C_C	1.139	1.080
Final vowels /_#	2.248	0.749
CONSONANTS	0.963	0.523
Consonant in onset /#C*_C*	0.372	0.177
Initial consonant /#_	0.345	0.151
Initial single consonant /#_V	0.371	0.152
Initial clustered consonant /#_C	0.192	0.077
Non-initial, onset consonant /#CC*_	0.283	0.159
Consonant in coda /C*_C*#	1.191	0.619
Non-final, coda consonant /_CC*#	0.199	0.105
Final consonant /_#	1.219	0.609
Final single consonant /V_#	1.231	0.658
Final clustered consonant /C_#	0.671	0.129

To summarize these results, vowels are more unpredictable than consonants (assuming a purely probabilistic writing algorithm). Both vowels and consonants are more uncertain when final. Consonants are more certain when in clusters, and especially so when at the beginning rather than at the end of a cluster, regardless of whether this is next to the vowel or word-marginal. It should be noted that a good deal of the uncertainty in final consonants results from my assigning final silent *e* as a polygraph with the preceding consonant rather than with the long-distance vowel. Those who prefer the opposite approach would transfer much of the uncertainty to medial vowels.

I also generated reports showing how the uncertainty of each phoneme differs before and after each possible phoneme, as well as word initially and word finally. This data is too long to include in this report, but may be accessed via ftp.[6] One interesting observation is that if the following phoneme is known, a phoneme can be spelt with uncertainty of 0.666 (0.371 over word tokens), a significant reduction over the 1.167 uncertainty of blindly spelling a phoneme. In the other direction, spelling a phoneme when the previous one is held constant, the uncertainty is a comparable 0.621 (0.251 with word tokens). Thus knowing the immediate context is valuable, but on the average, the left neighbour is as good as the right. Of course the details vary from phone to phone.

5 Rules

Although the uncertainty statistic H is an interesting and convenient measure of the number of possible spellings for a phone and their frequency distribution, it has important limitations, in that it assumes a probabilistic model of spelling. Therefore I have also constructed a rule-system for reducing the words in the monosyllable list to writing, given only accessible linguistic knowledge about the word.[7] I then attempted a crude measure of the complexity required to adequately handle the corpus, and compared it to the H statistic.

The rule system presented below is ideal, in the sense that it attempts to explain the regularity in the data, rather than the technique people actually use to spell. To be sure, I attempted to check the growth of the rule system by assigning to rules a "weight" equivalent to one exception, and in fact preferring exceptions to the equal number of rules. So a rule is not admitted unless it explicates at least two words more than the number of exceptions it robs from the productive output of some other rule; in the simplest case, a rule must explicate at least two forms. But it seems unlikely that rules are psychologically so light-weight. I was certainly surprised by the majority of the rules here presented. It also seems likely that a fairly unproductive rule may be tolerated if it is exceptionless, whereas rules with exceptions need to have a large ratio of output to exception.

In general, the rule system appeals to the phonetic environment of a phoneme. But in certain cases a great deal of complexity was avoided by appealing to the spelling. To avoid circularity in these cases, I permitted reference only to the spelling of previous phones in the word. This is because the assumption is that the phones are spelt in strict sequence. In particular, the multi-phone spellings such as *x* are treated only under the first phoneme, here /k/, with a notation showing that the rule also consumes the /s/.

Given that rules were considered to be about as heavy as exceptions, the weight of a rule system for a phoneme was determined by summing up the number of rules and exceptions. But a weight of ten obviously means one

[6] Files `before-report` and `after-report`, in the previously cited location.

[7] Appeal to etymology was ruled out, especially since people's intuitions about it, shaky at best, seem largely influenced by knowledge of the spelling.

thing if the phone is only found in ten words, and another if in a thousand, so I introduced a Complexity measure, which is the ratio of the weight to the number of times the phoneme appears in word types. The Complexity ratios discovered vary from 48.88% for /i/ to 1.27% for /p/ (all phones must have some Complexity, because they require at least one rule).

Details for the various phones are given below. For vowels, an average Complexity of 21.40% was found, ranging down to 1.81 for /a/. The Complexity measures have a 93% correlation with the H measures. The ranked list of complexities is:

/i/	48.88
/R/	47.89
/u/	47.30
/e/	31.96
/U/	31.67
/o/	31.58
/c/	27.44
/Y/	26.39
/W/	24.03
/E/	13.65
/C/	10.71
/V/	7.98
/y/	6.44
/A/	6.36
/O/	1.91
/a/	1.81

It is readily apparent that tense vowels and diphthongs have a higher Complexity than lax vowels; the weighted averages are 33.70% over against 6.87%. Since only tense vowels can be final, it might be wondered whether this difference is related to the fact that final vowels have higher H measures. In fact, tense vowels have approximately equal Complexity regardless of their position in the word. So it seems likely that, at least to some extent, the higher uncertainty of vowels in final position is due to the fact that they are all tense vowels.

For consonants the average weighted Complexity is 6.47%, and the correlation between H and Complexity is only 0.605422. One possible explanation for the low correlation is that for consonants subject to final doubling or silent *-e*, the variation is quite easily accounted for by rule, whereas most of the uncertainty of other consonants is not quite so tractable. At any rate, the rank order of the consonants, by rule Complexity, is as follows. It should be kept in mind that some phonemes with relatively simple rule systems, such as /Z/, tend to have a lot of Complexity because so few words have the sound; others, such as /z/, are complex because there are many rules and exceptions. The cases can easily be differentiated by inspecting the detailed presentation in the Appendix.

/Z/	42.86
/z/	30.11
/j/	24.42
/S/	12.82
/D/	12.44
/k/	12.06
/Q/	10.00
/s/	9.86
/T/	8.68
/f/	8.26
/n/	7.34
/g/	5.99
/m/	4.91
/w/	4.75
/l/	4.52
/r/	4.41
/v/	4.40
/h/	3.24
/d/	3.07
/t/	2.81
/H/	2.38
/b/	1.72
/q/	1.56
/p/	1.27

The average weighted Complexity for spelling any phoneme in particular was 11.19%.

The appendix contains a detailed presentation of the proposed rule system for spelling the English monosyllables. In addition to the rules and exceptions, under each phoneme is also given the measure of Complexity, and the synopsis as discussed above, including exact type and token counts for each spelling, and both type-based and token-based H measures. For an excellent summary of the historical factors leading to the present multiplicity of spellings for each phone, see [Cummings 1988], in particular the *Correspondences* chapters, which are organized by phoneme. The reader should keep in mind that the frequencies cited there are for words in general, not just monosyllables, and so are often radically different from the figures presented here.

6 Conclusions and Futures

The techniques developed in the course of this project are certainly useful for comparing the degree of regularity across different phones and/or across different environments in the word. The computation of the uncertainty statistic H is particularly easy to use, new analyses being performed by writing a short, largely stereotyped Lisp function. Its basis in statistical science is comforting, and to those who would rightly argue that it has no necessary relationship to *regularity* per se, since it appeals to no rule system, it can be answered that we have seen a moderate to strong correlation with rule complexity, as well as many examples (especially in the Appendix), where the statistic agrees with our intuitions about the difficulty of writing a phone by rule. Studies based on the H statistic could be used to investigate more closely just where the uncertainty in English spelling is. A practical application of this might be in paedagogy, where statistics could be used to locate the most regular spellings for early phonics training, and more irregular ones for later study. The use of the word token frequencies could be particularly useful here. Although this study has concentrated on a few environments of known importance in the paedagogical and psychological literature (e.g., clusters and the position within the syllable), one could also imagine a programme that would explore and report on particularly difficult environments, or pursue environments of decreasing uncertainty to report on contexts that might be of use in rules.

The work on a rule system for monosyllabic phoneme-oriented spelling had a few interesting results beyond the wealth of details. The most useful spelling rules were based on the values of adjacent phonemes – some phonemes caring more about the preceding, some the following; most caring about immediately adjacent phonemes, a few caring about two or more – or on the other spellings in the word. Natural phoneme classes were only rarely evidenced, with the exception of tense and lax vowels, which figured vitally in the spelling of following consonants. A general principle against mixing uniquely Greek spellings such as *ph* with uniquely English spellings such as *ea* proved disappointingly unprolific, its scope largely being subsumed by that of more productive rules. For future work, it would be interesting not only to attempt to discover potential rules as described in the paragraph above, but also to implement the rules in a programme, in order to check their validity and gather more statistical data. Unfortunately a real problem here is lack of clarity as to what constitutes a valid rule. As already mentioned, it is highly questionable whether writers actually maintain rules that generate as few as two words, or rules that have more exceptions than regular outputs. Psychological studies, involving for example requests that subjects spell phonotactically correct nonsense syllables, might help shed some light on which of these rules are psychologically real, and ultimately tell what properties rules must have to be realistic.

Although this study has clearly achieved a measure of success in exploring and documenting differences in regularity and uncertainty in the various parts of the phonemic subsystem of English spelling, it is less certain that it is now possible to give a simple answer to the question of how regular English orthography is, even for monosyllables. The answer that for an average phoneme the ratio of rules and exceptions to instances in word types (my Complexity ratio) is 11.19%, or that the uncertainty statistic H is 1.167 across word types and 0.789 across tokens, might not satisfy anyone very much. It might be more relevant if those statistics were compared to their counterparts for other alphabetically spelt languages, or if more confidence could be given to the psychological reality of the rules. As it stands, I can state that with 180 well-chosen rules, there remain 1,586 exceptions,[8] and any rejection of some of the rules on psychological criteria will only raise the number of exceptions. For spelling 4,724 core words of a subsystem with only nominal interference from other regular linguistic systems such as morphological analysis, that seems rather complex.

[8] These exceptions are per phoneme, not per word. It may of course be true that spellers tend to memorize whole words as simple gestalts, instead of potentially many exceptional phoneme correspondences for the same word. Of course if this is the case, and it applies to more regular spellings as well, the whole question of regularity in spelling takes on much added complexity.

References

[Chomsky 1970] Carol Chomsky. "Reading, Writing, and Phonology," *Harvard Educational Review* v. 40, pp. 287-310.

[Chomsky 1970?] Noam Chomsky. "Phonology and Reading," *Basic Studies in Reading,* edited by Levin and Williams. – New York : Harper and Row.

[Cummings 1988] D.W. Cummings. *American English Spelling : an informal description.* – Baltimore : Johns Hopkins Univ. Press, 1988.

[Francis 1982] W. Nelson Francis and Henry Kučera. *Frequency Analysis of English Usage : lexicon and grammar.* – Boston : Houghton Mifflin, 1982.

[Hanna 1966] Paul R. Hanna, et al. *Phoneme-Grapheme Correspondences as Cues to Spelling Improvement.* Washington : GPO, 1966.

[Random House 1987] *The Random House Dictionary of the English Language.* – 2nd ed. – New York : Random House, 1987.

[Sampson 1985] *Writing Systems.* Stanford, Cal. : Stanford Univ. Press, 1985.

[SCRL 1973] Speech Communications Research Laboratory. *American English Orthographic-Phonemic Dictionary.* Springfield, Va. : NTIS, 1973. Air Force Office of Scientific Research report AD-763 784.

[Venezky 1970] Richard L. Venezky. *The Structure of English Orthography.* – The Hague : Mouton, 1970.