# English orthography is not "close to optimal"

**Garrett Nicolai** and **Grzegorz Kondrak**
Department of Computing Science
University of Alberta
{nicolai,gkondrak}@ualberta.ca

## Abstract

In spite of the apparent irregularity of the English spelling system, Chomsky and Halle (1968) characterize it as "near optimal". We investigate this assertion using computational techniques and resources. We design an algorithm to generate word spellings that maximize both phonemic transparency and morphological consistency. Experimental results demonstrate that the constructed system is much closer to optimality than the traditional English orthography.

## 1 Introduction

English spelling is notorious for its irregularity. Kominek and Black (2006) estimate that it is about 3 times more complex than German, and 40 times more complex than Spanish. This is confirmed by lower accuracy of letter-to-phoneme systems on English (Bisani and Ney, 2008). A survey of English spelling (Carney, 1994) devotes 120 pages to describe phoneme-to-letter correspondences, and lists 226 letter-to-phoneme rules, almost all of which admit exceptions. Numerous proposals have been put forward for spelling reforms over the years, ranging from small changes affecting a limited set of words to complete overhauls based on novel writing scripts (Venezky, 1970).

In spite of the perceived irregularity of English spellings, Chomsky and Halle (1968) assert that they remarkably well reflect abstract underlying forms, from which the surface pronunciations are generated with "rules of great generality and wide applicability". They postulate two principles of an optimal orthographic system: (1) it should have "one representation for each lexical entry" (*consistency*); and, (2) "phonetic variation is not indicated where it is predictable by a general rule" (*predictability*). They conclude that "conventional orthography is [...] a near optimal system for the lexical representation of English words" (page 49), which we refer to as the *optimality claim*.

Chomsky and Halle's account of English orthography is not without its detractors. Steinberg (1973) argues against the idea that speakers store abstract underlying forms of separate morphemes and apply sequences of phonological rules during composition. Sampson (1985) cites the work of Yule (1978) in asserting that many common English word-forms provide counter-evidence to their vowel alternation observations. Derwing (1992) maintains that the observations only hold for five vowel alternations that can be predicted with simple spelling rules. According to Nunn (2006), the idea that spelling represents an abstract phonological level has been abandoned by most linguists. Sproat (2000) notes that few scholars of writing systems would agree with Chomsky and Halle, concluding that the evidence for a consistent morphological representation in English orthography is equivocal.

It is not our goal to formulate yet another proposal for reforming English orthography, nor even to argue that there is a need for such a reform. Furthermore, we refrain from taking into account other potential advantages of the traditional orthography, such as reflecting archaic pronunciation of native words, preserving the original spelling of loanwords, or maintaining orthographic similarity to cognates in other languages. Although these may be valid concerns, they are not considered as such by Chomsky and Halle. Instead, our primary objective is a deeper understanding of how the phono-

logical and morphological characteristics of English are reflected in its traditional orthography, which is currently the dominant medium of information exchange in the world.

In this paper, we investigate the issue of orthographic optimality from the computational perspective. We define metrics to quantify the degree of optimality of a spelling system in terms of phonemic transparency and morphological consistency. We design an algorithm to generate an orthography that maximizes both types of optimality, and implement it using computational tools and resources. We show experimentally that the traditional orthography is much further from optimality than our constructed system, which contradicts the claim of Chomsky and Halle.

## 2 Optimality

In this section, we define the notions of phonemic and morphemic optimality, and our general approach to quantifying them. We propose two theoretical orthographies that are phonemically and morphologically optimal, respectively. We argue that no orthographic system for English can be simultaneously optimal according to both criteria.

### 2.1 Phonemic optimality

A purely phonemic system would have a perfect one-to-one relationship between graphemes and phonemes. Rogers (2005) states that no standard writing system completely satisfies this property, although Finnish orthography comes remarkably close. For our purposes, we assume the International Phonetic Alphabet (IPA) transcription to be such an ideal system. For example, the IPA transcription of the word *viscosity* is [vɪskɑsəti]. We obtain the transcriptions from a digital dictionary that represents the General American pronunciation of English.

Phonemic transparency can be considered in two directions: from letters to phonemes, and vice versa. The pronunciation of Spanish words is recoverable from the spelling by applying a limited set of rules (Kominek and Black, 2006). However, there is some ambiguity in the opposite direction; for example, the phoneme [b] can be expressed with either 'b' or 'v'. As a result, it is not unusual for native Spanish speakers to make spelling mistakes. On the other hand, the orthography of Serbo-Croatian was originally created according to the rule "write as you speak", so that the spelling can be unambiguously produced from pronunciation. This does not mean that the pronunciation is completely predictable from spelling; for example, lexical stress is not marked (Sproat, 2000).

In this paper, we measure phonemic transparency by computing average perplexity between graphemes and phonemes. Roughly speaking, phonemic perplexity indicates how many different graphemes on average correspond to a single phoneme, while graphemic perplexity reflects the corresponding ambiguity of graphemes. We provide a formal definition in Section 5.

### 2.2 Morphological optimality

A purely morphemic writing system would have a unique graphemic representation for each morpheme. Chinese is usually given as an example of a near-morphemic writing system. In this paper, we construct an abstract morphemic spelling system for English by selecting a single alphabetic form for each morpheme, and simply concatenating them to make up words. For example, the morphemic spelling of *viscosity* could be 'viscous·ity'.[1]

We define morphemic optimality to correspond to the consistency principle of Chomsky and Halle. The rationale is that a unique spelling for each morpheme should allow related words to be readily identified in the mental lexicon. Sproat (2000) distinguishes between morpheme-oriented "deep" orthographies, like Russian, and phoneme-oriented "shallow" orthographies, like Serbo-Croatian.

We propose to measure morphemic consistency by computing the average edit distance between morpheme representations in different word-forms. The less variation morpheme spellings exhibit in a writing system, the higher the corresponding value of the morphemic transparency will be. We define the measure in Section 5.

It is impossible to achieve complete phonemic and morphemic optimality within one system designed for English spelling. For example, the stem morpheme of verb forms *hearing* and *heard* is

[1] Non-traditional spellings are written within single quotes. Morphemes may be explicitly separated by the centered dot character.

spelled identically but pronounced differently. If we changed the spellings to indicate the difference in pronunciation, we would move towards phonemic optimality, but away from morphemic optimality. Apart from purely phonographic or logographic variants, any English spelling system must be a compromise between phonemic and morphemic transparency. In this paper, we attempt to algorithmically create an orthography that simultaneously approaches the optimality along both dimensions.

## 3 Algorithm

In this section, we describe our algorithm for generating English spellings (Figure 1), which serves as a constructive proof that the traditional orthography is not optimal. Our objective is to find the best compromise between phonemic transparency and morphemic consistency. Section 3.1 explains how we derive a unique representation for each morpheme. Section 3.2 shows how the morpheme representations are combined into word spellings. Without a loss of generality, the generated spellings are composed of IPA symbols.

### 3.1 Morpheme representations

We start by identifying all morphemes in the lexicon, and associating each morpheme with sets of words that contain it (lines 1–3 in Figure 1). An example word set that corresponds to the morpheme *atom* is shown in Table 1. Words may belong to more than one set. For example, the word *atomic* will also be included in the word set that corresponds to the morpheme *-ic*. We make no distinction between bound and free morphemes.

As can be seen in Table 1, English morphemes often have multiple phonemic realizations. The objective of the second step (lines 4–11) is to follow the consistency principle by establishing a single representation of each morpheme. They suggest that orthographic representations should reflect the underlying forms of morphemes as much as possible. Unfortunately, underlying forms are not attested, and there is no commonly accepted algorithm to construct them. Instead, our algorithm attempts to establish a sequence of phonemes that is maximally similar to the attested surface allomorphs.

Table 1 shows an example of generating the common representation for a morpheme. We extract the phonemic representation of each allomorph in the word set, and perform a multi-alignment of the representations by pivoting on the longest representation of the morpheme (lines 5–8). For each position in the multi-alignment, we identify the set of phonemes corresponding to that position. If there is no variation within a position, we simply adopt the common phoneme. Otherwise, we choose the phoneme that is most preferred in a fixed hierarchy of phonemes. In this case, since [æ] and [ɑ] are preferred to [ə], the resulting morpheme representation is 'ætɑm'.

For selecting between variant phonemes, we follow a manually-constructed hierarchy of phonemes (Table 2), which roughly follows the principle of least effort. The assumption is that the phonemes requiring more articulatory effort to produce are more likely to represent the underlying phoneme. Within a single row, phonemes are listed in the order of preference. For example, alveolar fricatives like [s]

// Create word sets
1: **for each** word $w$ in lexicon $L$ **do**
2: **for each** morpheme $m$ in $w$ **do**
3: add $w$ to word set $S_m$
// Generate morpheme representations
4: **for each** word set $S_m$ **do**
5: $m_0 :=$ longest representation of $m$
6: **for each** word $w$ in $S_m$ **do**
7: $a_w :=$ alignment of $m_0$ and $w$
8: add $a_w$ to multi-alignment $A$
9: **for each** position i in $A$ **do**
10: select representative phoneme $r[i]$
11: $r_m := r[1..|m_0|]$
// Adopt a surface phoneme predictor
12: *Pronounce* := *Predictor (L)*
// Generate word representations
13: **for each** word $w = m_1 \ldots m_k$ **do**
14: $r := r_{m_1} \cdot \ldots \cdot r_{m_k}$
15: **for each** phoneme $r[i]$ in $r$ **do**
16: **if** *Pronounce*$(r[i]) \neq w[i]$ **then**
17: $r[i] := w[i]$
18: $r_w := r[1..|w|]$

Figure 1: Spelling generation algorithm. All representations consists of phonemes.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | æ | t | ə | m | | | | | *atom* |
| | | | æ | t | ə | m | z | | | | *atoms* |
| | | | ə | t | ɑ | m | ɪ | k | | | *atomic* |
| | | | ə | t | ɑ | m | ɪ | k | l | i | *atomically* |
| s | ʌ | b | ə | t | ɑ | m | ɪ | k | | | *subatomic* |
| | | | æ | t | ɑ | m | | | | | |

Table 1: Extracting the common morphemic representation .

are preferred to post-alveolar ones like [ʃ], in order to account for palatalization. Since our representations are not intended to represent actual underlying forms, the choice of a particular phoneme hierarchy affects only the shape of the generated word spellings.

### 3.2 Word representations

Ideally, polymorphemic words should be represented by a simple concatenation of the corresponding morpheme representations. However, for languages that are not purely concatenative, this approach may produce forms that are far from the phonemic realizations. For example, assuming that the words *deceive* and *deception* share a morpheme, a spelling 'deceive·ion' would fail to convey the actual pronunciation [dəsɛpʃən]. The predictability principle of Chomsky and Halle implies that phonetic variation should only be indicated where it is not predictable by general rules. Unfortunately, the task of establishing such a set of general rules, which we discuss in Section 7, is not at all straightforward. Instead, we assume the existence of an oracle (line 12 in Figure 1) which predicts the surface pronunciation of each phoneme found in the concatenation of the morphemic forms.

In our algorithm (lines 13–18), the default spelling of the word is composed of the representations of its constituent morphemes conjoined with a separator character. If the predicted pronunciation matches the actual surface phoneme, the "underlying" phoneme is preserved; otherwise, it is substituted by the surface phoneme. This modification helps to maintain the resulting word spellings reasonably close to the surface pronunciation.

For example, consider the word *sincerity*. Suppose that our algorithm derives the representations of the two underlying morphemes as 'sɪnsir' and 'ɪti'. If, given the input 'sɪnsir·ɪti', the predictor correctly generates the surface pronunciation [sɪnsɛrəti], we adopt the input as our final spelling. However, if the prediction is [sɪnsirəti] instead, our final spelling becomes 'sɪnsɛr·ɪti', in order to avoid a potentially misleading spelling. Since the second vowel was incorrectly predicted, we determine it to be unpredictable, and thus represent it with the surface phoneme, rather than the underlying one. The choice of the predictor affects only the details of the generated spellings.

| | |
|---|---|
| Stops | b d g p t k |
| Affricates | dʒ tʃ |
| Fricatives | ð v z ʒ θ f s ʃ h |
| Nasals | m n ŋ |
| Liquids | l r |
| Glides | j w |
| Diphthongs | aɪ ɔɪ aʊ |
| Tense vowels | i e o u ɑ |
| Lax vowels | æ ɛ ɔ ʊ ʌ |
| Reduced vowels | ɪ ə |
| deletion | _ |

Table 2: Hierarchy of phonemes.

## 4 Implementation

In this section, we describe the specific data and tools that we use in our implementation of the algorithm described in the previous section.

### 4.1 Data

For the implementation of our spelling generation algorithm, we require a lexicon that contains morphological segmentation of phonemic representations of words. Since we have been been unsuccessful in finding such a lexicon, we extract the necessary information from two different resources: the CELEX lexical database (Baayen et al., 1995), which includes morphological analysis of words, and the Combilex speech lexicon (Richmond et al., 2009), which contains high-quality phonemic transcriptions. After intersecting the lexicons, and pruning it of proper nouns, function words, duplicate forms, and multi-word entries, we are left with approximately 51,000 word-forms that are annotated both morphologically and phonemically.

In order to segment phonemic representations into constituent morphemes, we apply a high-precision phonetic aligner (Kondrak, 2000) to link letters and phonemes using the procedure described in (Dwyer and Kondrak, 2009). In rare cases where the phonetic aligner fails to produce an alignment, we back-off to alignment generated with *m2m-aligner* (Jiampojamarn et al., 2007), an unsupervised EM-based algorithm. We found that this approach worked better for our purposes than relying on the alignments provided in Combilex. We use the same approach to align variant phonemic representations of morphemes as described in Section 3.1.

The morphological information contained in CELEX is incomplete for our purposes, and requires further processing. For example, the word *amputate* is listed as monomorphemic, but in fact contains the suffix *-ate*. However, *amputee* is analyzed as

$$amputee = amputate - ate + ee.$$

This allows us to identify the stem as *amput*, which in turn implies the segmentations *amput·ee*, *amput·ate*, and *amput·at·ion*.

Another issue that requires special handling in CELEX involves recovering reduced geminate consonants. For example, the word *interrelate* is pronounced with a single [r] phoneme at the morpheme boundary. However, when segmenting the phoneme sequence, we need to include [r] both at the end of *inter-* and at the beginning of *relate*.

### 4.2 Predictor

The role of the predictor mentioned in Section 3.2 is performed by DIRECTL+ (Jiampojamarn et al., 2010), a publicly available discriminative string transducer. It takes as input a sequence of common morpheme representations, determined using the method described above, and produces the predicted word pronunciation. Since DIRECTL+ tends to make mistakes related to the unstressed vowel reduction phenomenon in English, we refrain from replacing the "underlying" phonemes with either [ə] or [ɪ].

An example derivation is shown in Table 3, where the *Underlying* string represents the input to DIRECTL+, *Predicted* is its output, *Surface* is the actual pronunciation, and *Respelling* is the spelling generated according to the algorithm in Figure 1.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Underlying: | foto | + | græf | + | ər | + | z |
| Predicted: | fotə | | græf | | ər | | z |
| Surface: | fətɑ | | grəf | | ər | | z |
| Respelling: | fotɑ | · | græf | · | ər | · | z |

Table 3: Deriving the spelling of the word *photographers*.

Since DIRECTL+ requires a training set, we split the lexicon into two equal-size parts with no morpheme overlap, and induce two separate models on each set. Then we apply each model as the predictor on the other half of the lexicon. This approach simulates the human ability to guess pronunciation from the spelling. Jiampojamarn et al. (2010) report that DIRECTL+ achieves approximately 90% word accuracy on the letter-to-phoneme conversion task on the CELEX data.

## 5 Evaluation measures

In this section, we define our measures of phonemic transparency and morphemic consistency.

### 5.1 Phonemic transparency

Kominek and Black (2006) measure the complexity of spelling systems by calculating the average perplexity of phoneme emissions for each letter. The total perplexity is the sum of each letter's perplexity weighted by its unigram probability. Since their focus is on the task of inducing text-to-speech rules, they also incorporate letter context into this definition. Thus, a system that is completely explained by a set of rules has a perplexity of 1.

The way we compute perplexity differs in several aspects. Whereas Kominek and Black (2006) calculate the perplexity of single letters, we take as units substrings derived from many-to-many alignment, with the length limited to two characters. Some letter bigrams, such as $ph$, $th$, and $ch$, are typically pronounced as a single phoneme, while the letter $x$ often corresponds to the phoneme bigram [ks]. By considering substrings we obtain a more realistic estimate of spelling perplexity.

We calculate the average orthographic perplexity using the standard formulation:

$$P_{ave} = \sum_c P_c e^{-\sum_i P_i log P_i} \quad (1)$$

| System | *viscous* | *viscosity* |
|---|---|---|
| T.O. | viscous | viscosity |
| IPA | vɪskəs | vɪskɑsəti |
| M-CAT | viscous | viscous·ity |
| ALG | vɪskɑs | vɪskɑs·ɪti |
| SR | viscous | viscosity |
| SS | viscus | viscosity |

Table 4: Example spellings according to various systems.

| System | Orth | Phon | Morph |
|---|---|---|---|
| T.O. | 2.32 | 2.10 | 96.11 |
| IPA | 1.00 | 1.00 | 93.94 |
| M-CAT | 2.51 | 2.36 | 100.00 |
| ALG | 1.33 | 1.72 | 98.90 |
| SR | 2.27 | 2.15 | 96.57 |
| SS | 1.60 | 1.72 | 94.72 |

Table 5: Orthographic, phonemic and morphemic optimality of spelling systems.

where $P_c$ is the probability of a grapheme substring in the dictionary, and $P_i$ is the probability that the grapheme substring is pronounced as the phoneme substring *i*. Note that this formulation is not contingent on any set of rules.

In a similar way, we compute the phonemic perplexity in the opposite direction, from phonemes to letters. The orthographic and the phonemic perplexity values quantify the transparency of a spelling system with respect to reading and writing, respectively.

### 5.2 Morphemic consistency

Little (2001) proposes to calculate the morphemic optimality of English spellings by computing the average percentage of "undisturbed letters" in the polymorphemic words with respect to the base form. For example, four of five letters of the base form *voice* are present in *voicing*, which translates into 80% optimal. The examples given in the paper allow us to interpret this measure as a function of edit distance normalized by the length of the base form.

We make three modifications to the original method. First, we compute the average over all words in the lexicon rather than over word sets, which would give disproportionate weight to words in smaller word sets. Second, we normalize edit distance by the number of phonemes in a word, rather than by the number of letters in a spelling, in order to avoid penalizing systems that use shorter spellings. Finally, we consider edit operations to apply to substrings aligned to substrings of phonemes, rather than to individual symbols. In this way, the maximum number of edit operations is equal to the number of phonemes. The modified measure yields a score between 0 and 100%, with the latter value representing morphemic optimality.

As an example, consider the word set consisting of six word-forms: *snip*, *snips*, *snipped*, *snipping*, *snippet*, and *snippets*. The first two words, which represent the base morpheme as *snip*, receive a perfect score of 1 for morphemic consistency. The remaining four words, which have the morpheme as *snipp*, obtain the score of 75% because one of the four phonemes is spelled differently from the base form. For free morphemes, the base form is simply the spelling of the morpheme, but for bound morphemes, we take the majority spelling of the morpheme.

## 6 Quantitative comparison

We compare the traditional English orthography (T.O.) to three hypothetical systems: phonemic transcription (IPA), morpheme concatenation (M-CAT), and the orthography generated by the algorithm described in Section 3 (ALG). In addition, we consider two proposals submitted to the English Spelling Society: a minimalist spelling reform (SR) of Gibbs (1984), and the more comprehensive SoundSpel (SS) of Rondthaler and Edward (1986). Table 4 lists the spellings of the words *viscous* and *viscosity* in various orthographies.

Table 5 shows the values of orthographic and phonemic transparency, as well as morphemic consistency for the evaluated spelling systems. By definition, phonemic transcription obtains the optimal transparency scores of 1, while simple morphological concatenation receives a perfect 100% in terms of morphemic consistency.

The results in Table 5 indicate that traditional orthography scores poorly according to all three measures. Its low orthographic and phonemic transparency is to be expected, but its low morphemic

| Rule | Input | Output |
|---|---|---|
| e-deletion | voic**e**·ing | voicing |
| y-replacement | industr**y**·al | industr**i**al |
| k-insertion | panic·ing | panic**k**ing |
| e-insertion | church·s | church**e**s |
| consonant doubling | get·ing | get**t**ing |
| f-voicing | kni**f**e·s | kni**v**es |

Table 6: Common English spelling rules with examples.

| Rule | Writing | Reading |
|---|---|---|
| e-deletion | 98.8 | 67.1 |
| y-replacement | 93.5 | 95.8 |
| k-insertion | 100.0 | 1.0 |
| e-insertion | 100.0 | 98.7 |
| consonant doubling | 96.3 | 36.3 |
| f-voicing | 33.3 | 14.7 |

Table 7: Applicability of common spelling rules.

consistency is striking. Traditional orthography is not only far from optimality, but overall seems no more optimal than any other of the evaluated systems.

Searching for the explanation of this surprising result, we find that much of the morphemic score deduction can be attributed to small changes like dropping of the silent *e*, as in 'make' + 'ing' = 'making'. These types of inconsistencies counter-weigh the high marks that traditional orthography gets for maintaining consistent spelling in spite of unstressed vowel reductions.

The prevalence of silent $e$'s in traditional orthography undeniably diminishes its morphemic consistency. Nor is the device necessary to represent the pronunciation of the preceding vowel; for example, SoundSpel has those words as 'maek' and 'maeking'. However, one can argue that such minor alterations should not be penalized because English speakers subconsciously take them into account while reading. In the next section, we describe an experiment in which we pre-process words with such orthographic rules, in order to determine how much they influence the optimality picture.

## 7 Spelling rules

Table 6 lists six common English spelling rules that affect letters at morpheme boundaries, of which the first five are included in the textbook account of Jurafsky and Martin (2009, page 63). We conducted an experiment to determine the applicability of these rules by computing how often they fired when triggered by the correct environment.[2] We tested the rules in both directions, with respect to both writing and reading applicability. Writing rules are applied to morphemes when they are in the correct environment. For example, the *k-insertion* rule fires if the morpheme ends in a $c$ and the next morpheme begins with $e$ or $i$, as in *panic·ing*. On the other hand, reading may involve recovering the morphemes from the surface forms. For example, if the stem ends in a $tt$ and the affix begins with an $i$, the *consonant doubling* rule implies that the free form of the morpheme ends in a single $t$, as in *getting*.

The results in Table 7 show that the rules, with the exception of the *f-voicing* rule, have high applicability in writing. Most rules, however, cannot be trusted to recover the morpheme spellings from the surface form. For example, following the consonant doubling rule would cause the reader to incorrectly infer from the word *butted* that the spelling of the verb is *but*. This is significant considering that Chomsky and Halle define orthography as a system for *readers* (page 49).

Notwithstanding the unreliability of the spelling rules, we incorporate them into the computation of the morphemic consistency of the traditional orthography. We apply the rules from a reading perspective, but assume some morphemic knowledge of a reader. Whereas we consider a rule to misfire if it does not apply in the correct environment when calculating applicability, as in Table 7, when calculating morphemic consistency, we allow the rules to be more flexible. We consider a morpheme to match the prototype if either the observed form or the form modified by the spelling rule matches the prototype.

## 8 Discussion

Figure 2 shows a two-dimensional plot of orthographic perplexity vs. morphemic consistency. The (unattainable) optimality is represented by the lower

[2]The conditioning environments of the rules were implemented according to the guidelines provided at *http://www.phonicslessons.co.uk/englishspellingrules.html*.

left corner of the plot. The effect of accommodating the spelling rules within the traditional orthography is illustrated by an arrow, which indicates an increase in morphemic consistency from 96.11 to 98.90.

The ALG(L) system represents a version of the ALG system in which the IPA symbols are respelled using combinations of the 26 letters of the Roman alphabet, with the morpheme boundary symbol removed. This change, which is intended to make the comparison with the traditional orthography more interpretable, increases the orthographic perplexity from 1.33 to 1.58. Furthermore, we ensure that ALG(L) contains no homographs (which consitute 2.6% of the lexicon in ALG) by reverting to a traditional spelling of a morpheme if necessary. Since the respelling applies to all instances of that morpheme, it has no effect on the morphemic consistency, but results in a small increase of the orthographic perplexity to 1.61.

The plot in Figure 2 shows that, even after accounting for the orthographic rules, traditional orthography does not surpass the level of morphemic consistency of ALG. With the same writing script and no homographs, ALG(L) is less than half the distance from the orthographic optimality. On the other hand, neither of the spelling reform proposals is substantially better overall than the traditional orthography.

Inspection of the spellings generated by our algorithm reveals that it generally maintains consistent spellings of morphemes. In fact, it only makes a change from the underlying form in 3660 cases, or 7.2% of the words in the dictionary. Consider the morpheme *transcribe*, which is traditionally spelled as 'transcrip' in *transcription*. Even if we disregard the final 'e' by invoking the *e-deletion* spelling rule, the morphemic consistency in the traditional orthography is still violated by the 'b'/'p' alternation. Our predictor, however, considers this a predictable devoicing assimilation change, which occurs in a number of words, including *subscription* and *absorption*. Consequently, the spellings generated by the algorithm preserve the morpheme's 'b' ending in all words that contain it. In addition, the algorithm avoids spurious idiosyncrasies such as *four/forty*, which abound in traditional orthography.

The spellings generated by the algorithm are also

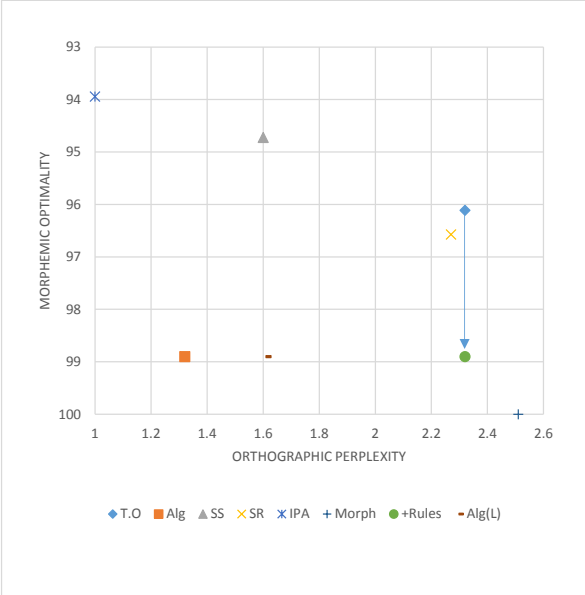


Figure 2: Morphemic and orthographic optimality of various spelling systems.

much more phonemically transparent, particularly for vowels. Phonemically, ALG(L) improves on the traditional orthography mostly by making the spelling more predictable, For example, 'a' represents the phoneme [æ] in 91.7% of the cases in the generated spellings, as opposed to only 36.5% in traditional orthography.

## 9 Conclusion

We have analyzed English orthography in terms of morphemic consistency and phonemic transparency. According to the strict interpretation of morphemic consistency, traditional orthography is closer to the level of a phonemic transcription than to that of a morphemic concatenation. Even if orthographic rules are assumed to operate cost-free as a preprocessing step, the orthographic perplexity of traditional orthography remains high.

While phonemic transparency and morphemic consistency are at odds with each other, we have provided a constructive proof that it is possible to create a spelling system for English that it is substantially closer to theoretical optimality than the traditional orthography, even when it is constrained by the traditional character set. This contradicts the claim that English orthography is near optimal.

## Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada, and the Alberta Innovates – Technology Futures.

## References

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

Edward Carney. 1994. *A Survey of English Spelling*. Routledge.

Noam Chomsky and Morris Halle. 1968. The sound pattern of English.

Bruce L Derwing. 1992. Orthographic aspects of linguistic competence. *The linguistics of literacy*, pages 193–210.

Kenneth Dwyer and Grzegorz Kondrak. 2009. Reducing the annotation effort for letter-to-phoneme conversion. In *Proceedings of ACL-IJCNLP*, pages 127–135.

Stanley Gibbs. 1984. The Simplified Spelling Society's 1984 proposals. *Journal of the Simplified Spelling Society*, 2:32.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York, April. Association for Computational Linguistics.

Sitichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2010. Integrating Joint n-gram Features into a Discriminative Training Framework. In *Proceedings of NAACL-2010*, Los Angeles, CA, June. Association for Computational Linguistics.

Dan Jurafsky and James H Martin. 2009. *Speech & language processing*. Pearson Education India, 2nd edition.

John Kominek and Alan W. Black. 2006. Learning pronunciation dictionaries: Language complexity and word selection strategies. In *HLT-NAACL*, pages 232–239.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of NAACL 2000: 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 288–295.

Joseph R Little. 2001. The optimality of English spelling.

Anneke Marijke Nunn. 2006. *Dutch orthography: A systematic investigation of the spelling of Dutch words*. The Hague: Holland Academic Graphics.

Korin Richmond, Robert AJ Clark, and Susan Fitt. 2009. Robust LTS rules with the Combilex speech technology lexicon. pages 1295–1298, September.

Henry Rogers. 2005. *Writing Systems*. Blackwell.

Edward Rondthaler and J LIAS Edward. 1986. Dictionary of simplified American Spelling.

Geoffrey Sampson. 1985. *Writing systems: A linguistic introduction*. Stanford University Press.

Richard Sproat. 2000. *A computational Theory of Writing Systems*. Cambridge.

Danny D Steinberg. 1973. Phonology, reading, and Chomsky and Halle's optimal orthography. *Journal of Psycholinguistic Research*, 2(3):239–258.

Richard L Venezky. 1970. *The structure of English orthography*, volume 82. Walter de Gruyter.

Valerie Yule. 1978. Is there evidence for Chomsky's interpretation of English spelling? *Spelling Progress Bulletin*, 18(4):10–12.