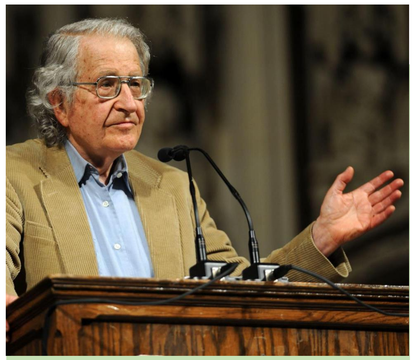
English Orthography is not “close to optimal”

Garrett Nicolai and Greg Kondrak

UNIVERSITY OF ALBERTA



Noam Chomsky

Chomsky and Halle (1968) claim that

- *conventional orthography is [. . .] a near optimal system for the lexical representation of English words.*

Further, an optimal orthography:

- *should have one representation for each lexical item*
- *phonetic variation is not indicated where it is predictable by a general rule*

We define optimality along two criteria:

- phonemic transparency, and
- morphological consistency

We show that traditional English orthography is not “close to optimal”.

Morris Halle

System	“viscous”	“viscosity”
Traditional Orthography	viscous	viscosity
Phonemic Representation	vɪskəs	vɪskəsəti
Morphological Concatenation	viscous	viscousity
Algorithmic Generation	vɪskəs	vɪskəs·ɪti
Spelling Reform	viscous	viscosity
SoundSpel	viscus	viscosity

System	Orthographic Perplexity	Phonemic Perplexity	Morphological Consistency
Traditional Orthography	2.32	2.10	96.11
Phonemic Representation	1.00	1.00	93.94
Morphological Concatenation	2.51	2.36	100.00
Algorithmic Generation	1.33	1.72	98.90
Spelling Reform	2.27	2.15	96.50
SoundSpel	1.60	1.72	94.72

Phonemic Optimality

$$P_{ave} = \sum_{i}^{C} P_c e^{-\sum P_i log P_i}$$

Orthographic Perplexity of “a”		
Phoneme	Example	Prob.
/ɑ/	fall	0.07
/ə/	balloon	0.33
/e/	safe	0.23
/æ/	match	0.32
Perplexity		3.51

Phonemic Perplexity of /æ/		
Spelling	Example	Prob.
“a”	match	0.998
“au”	laugh	0.001
“ai”	plaid	0.001
Perplexity		1.01

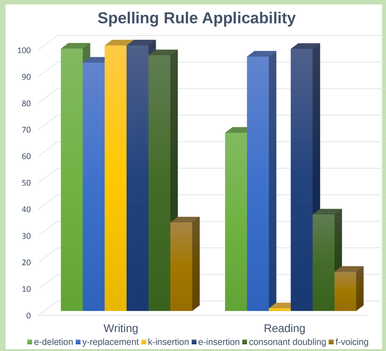
Morphological Optimality

Morphological Optimality is the average modified, normalized edit distance across all morphemes.

Spelling	Morph. Spelling	Pronunciation	Aligned	Consistency
Prototype = snip		/snɪp/	s:s n:n iːɪ p:p	4/4
snips	snip	/snɪp/	s:s n:n iːɪ p:p	4/4
snipping	snipp	/snɪp/	s:s n:n iːɪ pp:p	3/4
snipped	snipp	/snɪp/	s:s n:n iːɪ pp:p	3/4
snippet	snipp	/snɪp/	s:s n:n iːɪ pp:p	3/4
snippets	snipp	/snɪp/	s:s n:n iːɪ pp:p	3/4
Morphological Consistency				20/24

Spelling Rules

Rule	Example	
e-deletion	voice·ing	voicing
y-replacement	industry·al	industrial
k-insertion	panic·ing	panicking
e-insertion	church·s	churches
consonant doubling	get·ing	getting
f-voicing	knife·s	knives



Alphabet and Homographs

- To remove any advantage our system might gain from an expanded character set, we respell using Latin characters.

Alphabetic respelling of “pier”	
Traditional Spelling	pier
Recommended Respelling	*pir*
Alphabetic Respelling	*peer*

- We resolve homographs by respelling homographic morphemes with their traditional spelling.

Homographic respelling of /pir/			
Traditional Spelling	peer	pier	piers
Alphabetic Respelling	*peer*	*peer*	*peerz*
Homographic Respelling	*peer*	*pier*	*pierz*

Spelling Generation Algorithm

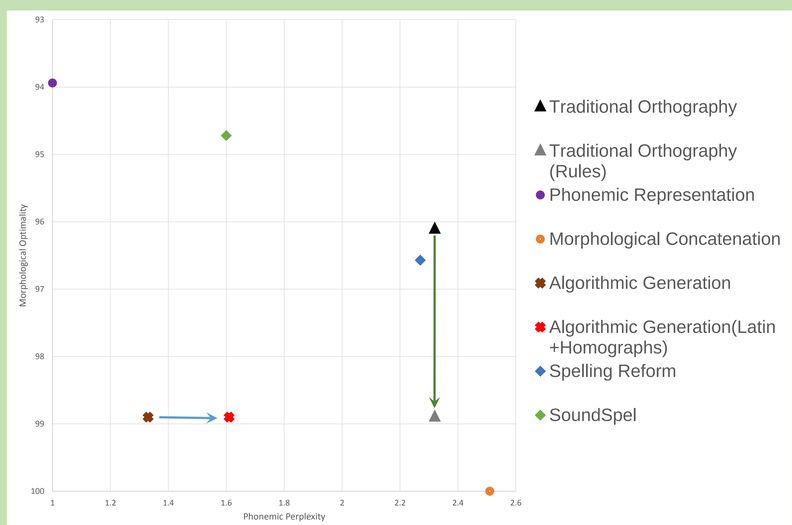
```
//   Create word Sets
1:  for each word w in lexicon L do
2:      for each morpheme m in w do
3:          add w to word set S_m

//   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 ... m_k do
14:     r := r_m1 · ... · r_mk
15:     for each phoneme r[i] in r do
16:         if Pronounce(r[i]) w[i] then
17:             r[i] := w[i]
18:     rw := r[1..|w|]
```

black	*berry*	*sincere*	*in-*	*-able*
“black”	“berry”	“sincere”	“insincerity”	“inescapable”
“blacker”	“strawberry”	“sincerity”	“inescapable”	“erasable”
…	…	…	…	…
“blackberry”	“blackberry”	“insincere”	“inevitable”	“inevitable”

Aligned Pronunciation	Spelling
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”
ə t ɑ m ɪ k	“subatomic”
Morpheme Representation:	*ætam*

Underlying:	f o t o · g r æ f · ə r · z
Predicted:	f o t ə · g r æ f · ə r · z
Surface:	f ə t ɑ · g r ə f · ə r · z
Respelling	f o t ɑ · g r æ f · ə r · z

Phonemic Hierarchy	
Class	Phonemes
Stops	/ b d g p t k /
Affricates	/ʤ ʧ/
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	_



Conclusions

- 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.
- 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, contradicting the claim that English orthography is near optimal.