

Spelling Reform: An Engineering Approach

Kamran Ossia (kamran@euspell.org)

Introducing Euspell, a conservative spelling reform system which includes: a 205000-word lexicon where each word is tagged with its common parts of speech, a 35000-word PLS lexicon of new spellings with their IPA pronunciations for use in speech synthesis, and a 46000-word lexicon listing the parts of speech of the new spellings. Also provided are browser extensions and word processor add-ins, PDF and epub readers, and in-browser dictation. The new spellings are chosen to remain within the spirit of English pronunciation and to make human and machine comprehension easier. No diacritics or new letters are introduced, and minimal changes are made to scientific and technical terms. Most new spellings involve removing unneeded letters, changing the verb ending *-s* to *-z* for third person singular, and *-ed* to *-d* for past tenses when the *e* is silent, all designed to improve consistency and to make reading easier. Two new forms *qh* and *uw* are introduced to disambiguate the *ch* and *ow* sounds. Source code is released under GPL-3 and lexicons are released under CC BY-SA 4.0, available at github.com/ossia/.

Nomenclature

The Main Lexicon: A 205000-entry csv file containing English words, their parts of speech, three-digit codes indicating the reason for spelling change, and the new spellings, if any.

PoS: Part of speech, e.g. noun, verb, adjective.

CLAWS-7: University of Lancaster PoS tagset [1].

NLP: Natural language processing. For our purpose, NLP involves mainly text classification.

Grapheme: The smallest unit in a system of writing a language that can express a difference in sound or meaning.

Homophone: Words that sound the same but have different spellings, e.g. *so* and *sew*.

Homograph: Words that have the same spelling but different sounds or meanings, e.g. *live* and *bears*, some of which may split a homograph into two or more words, e.g. *tear* → *tear* / *taer*.

Spurious letter: one whose removal or replacement brings the word more in line with common English spelling rules, e.g. the final *e* in *have*, *give*, *examine* (→ *hav*, *giv*, *examin*), the *o* in *lose*, *move*, *flood* (→ *luse*, *moov*, *fludd*).

Disambiguation: Changing the spelling of a word so that its meaning, PoS, or pronunciation is not ambiguous, either by removing or replacing spurious letters, or by changing grammatical endings.

PLS: Pronunciation Lexicon Specification, an XML (extensible markup language) file describing the phonetic IPA (International Phonetic Association) pronunciation of each word.

Introduction

English became the international language of science, commerce through a unique combination of attributes: expressivity with a vast and expandable vocabulary, a simple grammar free from gender and case endings, the use of the Latin alphabet without diacritics, a comprehensive set of verb tenses, absence of diglossia, and a systematic way of generating new technical and scientific terms. The main drawback of English is its inconsistent spelling. We believe that we have found the sweet spot of maintaining readability, improving comprehension, and reducing spelling inconsistencies and ambiguities by creating an entire ecosystem for the new spelling, including spelling and grammar check lexicons for word processors, browser extensions to convert web pages in real time, plugins for text-to-speech and dictation apps, and PDF/epub/text converters for desktop and mobile platforms.

History

The English language does not have a ‘regulatory agency’ similar to the French Academy or the Council for German Orthography. Spelling changes in English generally happen through mass adoption. While Old English was largely phonetic when it first adopted the Latin alphabet, linguistic evolution and the arrival of Norman scribes in the 11th century made its orthography highly irregular. This inconsistency was worsened by distinct regional dialects and the misguided ‘Latinization’ of certain words. The invention of the printing press during the Great Vowel Shift exacerbated the situation. Spellings were chosen almost at random from existing variations. Additionally, early printers added or removed letters simply to justify text margins rather than adjusting the spacing, which further delayed the establishment of a consistent spelling system.

One of the early attempts at standardizing English spelling came from Richard Mulcaster [4]. Taking a more moderate approach, he sought to standardize existing spellings, establishing conventions like using a silent *e* to indicate long vowels. Many attempts over centuries have had mixed results. On one extreme was George Bernard Shaw who invented an entirely new alphabet [5], while others have promoted phonetic spelling or adding diacritical marks. The Americanized spelling promoted by Noah Webster [6] made lasting changes including *-ise* → *-ize* for Greek-derived verbs (which disambiguates a few words, e.g. *analyses*) and dropping unneeded letters as in *colour* → *color*. The majority of Euspell changes are a continuation of this approach, as explained in the guidelines below. In the early 1900’s, Andrew Carnegie financed efforts to create a simplified spelling system, published by Paine [7]. Although quite conservative in its reform, it included a large set of rules which were not always consistent, plus a set of case-by-case changes. Many of Euspell changes coincide with those of Paine. Upward [8] has proposed ‘Cut Spelling’ where redundant letters and many unstressed vowels are removed. However, there are inconsistencies in their new spelling, e.g. *burlesque* becomes *burlesq* but *brusque* remains unchanged. Disambiguating diatonics is done for *-ate* forms but not for the *ow* and *ch* graphemes or the *-s* ending. Collisions are even introduced, e.g. *can’t* → *cant*, *guild* → *gild*, *write* → *rite*, *weather/whether* → *wethr*, etc. Some pronunciations are also changed, e.g. *bear* → *ber*, *could* → *cud*, and *putt* → *put*. Proper nouns and names of people and places are also respelled, creating potential legal and administrative issues.

Motivation

The cost of inconsistent spelling falls on two groups. For human readers, the burden is heaviest on learners and on those with reading difficulties. Memorizing irregular spellings on a case-by-case basis requires a different level of effort than remembering consistent rules such as when a vowel is long or short. The size of that difference has been measured. Seymour et al. [9] tested foundation literacy in thirteen European orthographies using the same three tasks in each: letter knowledge, reading familiar high-frequency words, and reading simple nonwords. At the end of the first school year, children reading Finnish scored 98% on familiar words, Greek 98%, German 98%, Italian 95%, Spanish 95%. Children reading English scored 34%. On nonwords — the test of decoding proper, uncontaminated by whole-word memory — the same children scored 29%, against 95% for Finnish and 94% for German. Reading speed showed the same break: most languages reached fluency at under 2 seconds per word within the first year, while the English-language sample took 7.8 seconds per word and was classified as an outlier. The authors' summary is that “the rate of development in English is more than twice as slow as in the shallow orthographies”, and that 7.5 years or more of reading is needed before English readers match the European levels reached in the first school year.

The study points out three findings. Letter-sound knowledge was not the problem: the English-language sample scored 94% on letters, comparable to every other group, so the deficit is in the orthography rather than in the alphabet or in how it was taught. The comparison used the higher-socioeconomic sample, whose reading ages were already ahead of expectation. The effect persists into the second school year, where the same children reached 76% on familiar words — still below what every shallow orthography achieves a year earlier. Seymour et al. attribute the delay to two independent causes: syllabic complexity, which affects decoding, and orthographic depth, which affects both word and nonword reading. English is unusual in being extreme on both. Respelling addresses only the second. A reform of the orthography cannot make English syllables simpler, and no claim is made here that it would close the whole gap, only that it removes one of the two documented causes, and the one that is amenable to change.

The same pattern appears in the clinical literature. Landerl et al. [10] found that German and English dyslexic children share the same underlying phonological deficit but differ substantially in how severely it shows up in reading, the German children reading nonwords far more accurately than their English counterparts. Paulesu and colleagues [11] reached the same conclusion from neuroimaging across Italian, French and English readers: one biological condition, expressed more or less disablingly according to the depth of the orthography the reader's native language.

The KPMG Foundation [12] estimated that the pupils leaving primary school in England and Wales each year unable to read and write properly — 6% of eleven-year-olds, about 38700 children — will each cost the public purse between £44797 and £53098 by the age of 37, through the associated effects on education, health, employment and crime: between £1.73 billion and £2.05 billion for every year's cohort. That figure is for one jurisdiction and covers literacy failure from all causes, not orthography alone.

For machines, the same irregularity is a direct obstacle: text-to-speech engines mispronounce homographs, and PoS taggers must guess the tags where the orthography gives no signal. A wider collection of research on the subject is maintained by the English Spelling Society [14].

The reason this has remained hard is that the interesting cases are not simple substitutions. Whether *read* rhymes with *reed* or *red*, whether the *-s* in a word marks a plural or a third-person verb, none of these can be settled by a context-free rule, yet each demands a different spelling if the reform is to reduce ambiguity. A reform that ignores such cases leaves the hardest part of the problem untouched.

Euspell is therefore built as an engineering system rather than a set of prescriptions. Conversion is performed by a deterministic pipeline over a lexicon, so the same input always yields the same output with no human judgment at conversion time; the genuinely ambiguous cases are resolved by classification over local context, treating disambiguation as the machine-learning problem it is. Because the reform is expressed as software, it can be delivered where reading and writing actually happen: browsers, document readers, word processors, and text-to-speech and dictation, and applied reversibly, so that a reader loses nothing by trying it and can recover the original text.

The Lexicons

euspell_lexicon.csv is a comprehensive lexicon of English with about 205000 entries in comma-separated value format with four fields: Word (the original English word), PoS (all common parts of speech for each word), Encoding (a three-digit number indicating if and how the spelling will be changed, see Appendix D), and euspending (the new spelling(s) of the word, if any). All different parts of speech for nouns (singular, plural, or uncountable), four verb forms, adjectives, and adverbs, etc. are included in the Lexicon. The PoS tags are adapted from CLAWS7 [1] which is the most granular tagset with 138 tags (See Appendix B). The following lines are typical:

```
barrelled,JJ|VVD|VFN,021,barreld
barrelling,JJ|VVG,631,barreling
barrels,NN2|VVZ,012,barrels|barrelz
barren,JJ,000,[]
```

euspell_tts.pls. This XML pronunciation lexicon specification file contains over 35000 words with new spellings and IPA pronunciations [3] (see Appendix C), to be used by speech synthesis software. The following lines are typical:

```
<lexeme><grapheme>buwz</grapheme><phoneme>baʊz</phoneme></lexeme> <!-- bows -->
<lexeme><grapheme>buyz</grapheme><phoneme>baɪz</phoneme></lexeme> <!-- buys -->
<lexeme><grapheme>buzzd</grapheme><phoneme>bʌzd</phoneme></lexeme> <!-- buzzed -->
<lexeme><grapheme>buzzez</grapheme><phoneme>'bʌzɪz</phoneme></lexeme> <!-- buzzes -->
```

euspell_pos.tsv. This tab-separated file with three fields (word, lemma, PoS) is a list of over 46000 entries for the new spellings in the Penn Treebank (PTB) [2] format required by grammar checkers. Each word has as many entries as its parts of speech, for example the word *barreld* has three entries:

barreld	barreld	JJ
barreld	barreld	VBD
barreld	barreld	VDN

Rules

1. No diacritical marks or new apostrophes will be added.
2. The new spelling should not coincide with an existing one.
3. The third-person singular present tense ending *-s* will be changed to *-z*.
4. The past tense verb ending will be *-d* unless the *e* is needed for correct pronunciation.
5. The silent *gh* will be reduced to *h*, with the preceding vowels adjusted as needed.
6. The *f*-sounding *gh* will be changed to *ff*, with the preceding vowels adjusted as needed.
7. The string *ch* will be changed to *qh* only when pronounced **tʃ** as in *church*.
8. The string *ow* will be changed to *uw* only when pronounced **aʊ** as in *cow*.
9. The French strings *ch* and *oi* pronounced *sh* and *wa* will be spelled as such.
10. The string *ou* pronounced **u:** will be replaced with *u* or *oo* when appropriate.
11. Spurious letters will be removed or replaced to disambiguate meaning or pronunciation.
12. Capitalized common nouns and adjectives may be changed; otherwise proper nouns and names of people and places will be left alone. Transliterated foreign names may be respelled.

Guidelines

The main goal of Euspell is for an English speaker to be able to read the text with relative ease; Euspell words are in almost all cases readily recognizable by the reader, differing by only one letter.

Familiarity. The new text must be close enough to traditional text that reader familiar with either spelling should be able to read the other text comfortably. This is the case for all shortened words. The words with the new forms *qh* and *uw* (described below) may take a few days to get used to.

Disambiguation. Euspell attempts to disambiguate homographs as much as practical, improving both human and machine comprehension. The largest category belongs to the *-s* ending, retained for plural nouns and changed to *-z* for third-person-singular verbs. The NN2|VVZ disambiguation algorithm is based on support vector machine (SVM) classification and has an accuracy of 94%. It struggles with minimalist headlines and words with insufficient context. The next largest category is JJ|VV0 words ending in *-ate*, such as *graduate*, where the JJ/NN1 form will lose the final *e*. A three-way disambiguation involves the words *leads*, *showers*, and *winds*, and the variants of *slough* (except *sloughs*). A four-way disambiguation involves the words *bows*, *fillets*, *rows*, *sloughs*, and *tears*.

Consistency. Another goal of Euspell is removing or replacing spurious letters which violate the spirit of English spelling: *live* (verb) should be spelled *liv*, *examine* should be *examin*, and so forth, because the spurious final *e* signals a long preceding vowel which is not the case here. Nearly 2000 words have their final *e* dropped to conform, including adjectives ending in *-ive* and *-ate*, e.g. *activ* and

fortunat. About 150 words ending in *-ate[s]* have two pronunciations, e.g. *graduate* is **grædʒuert** as verb and **grædʒurt** as noun/adjective, thus splitting into *graduate/graduāt*.

Avoiding collisions. Creating collisions with existing words goes against the philosophy of Euspell. Of the nearly 35000 words with new spellings, only a few have resulted in collisions with existing words, e.g. *heart* → *hart* (male red deer) and *weather* → *wether* (neutered ram). These words can be distinguished from context. Until a solution is found to distinguish the two short *u* sounds **ʊ** and **ʌ** (*put* versus *putt*), the words *could*, *crowed*, *should*, and *would* will be spelled *coodd*, *crowdd*, *shoodd*, and *woodd*. Without the doubled *d*'s these would result in collisions. Other past tenses avoiding collision are words like *barred*, *gelled*, *mannd* which will be *barrd*, *gelld*, and *mannd* respectively. Euspell standardizes on American spelling, so technically the converted British spellings result in collisions. A set of ‘good’ collisions are words reverting to their former spellings, as shown in Table 1.

Table 1. Words reverting to former spellings

Current	Old & New	Current	Old & New	Current	Old & New
aghaſt	agast	debt	dett	indict	indite
anchor	ancor	doubt	dout	learn	lern
aweless	awless	forfeit	forfet	ſcythe	ſythe
boulder	bowlder	guard	gard	ſurfeit	ſurfet
build	bild	harangue	harang	tongue	tung
brusque	brusk	haulm	haum	victual	vittle
counterfeit	counterfet	hearken	harken	wealth	welth
curtain	curtin	hearse	herse	wind (waſnd)	wynd

Reversibility. The rarity of collisions means that Euspell is reversible, i.e. the original document can be recovered, albeit with American spelling. **Note: Do not convert this paper or other documents with mixed traditional and Euspell content.**

Treatment of silent letters. Silent letters fall into a few categories. In some cases such as *hymn* and *bomb*, pronouncing them requires additional effort, so they tend to be dropped in normal speech, but they can appear in derived forms such as *hymnal* and *bombardment* (but not *bombing*). These silent letters are left alone. Another category is where silent letters indicate a preceding long vowel, as in *rogue*, which are dealt with on a case-by-case basis to follow the relevant rules. Spurious silent letters as the *ue* in *colleague* are removed.

Disambiguation of the -s ending. The largest single change in Euspell is to distinguish the *-s* ending of plural nouns from third-person-singular verbs (VVZ) by changing the latter ending to *-z*. This single change will be of great help in human and machine comprehension. Although the *s* in *is* is not an ending, it will also be changed to *-z* to be consistent with other VVZ forms and to distinguish it from the possessive ending *-’s*: *Jim’s home* (the house of Jim) and *Jim’z home* (Jim has returned).

It's not it's any more. The confusion of *its* and *it's* has become widespread in recent years, seen even in advertising and published articles. The word *it's*, meaning *it is*, will no longer exist in the new spelling, replaced with *it'z*. *Its* will remain unchanged.

Past tense ending. It has been argued that adding *-ed* as the past tense ending is a simple rule that should not be changed. That is not entirely accurate, as additional rules exist to double the preceding consonant in some cases, and adding only *-d* in other cases. It may even create ambiguities such as *secreted* (past tense of both *secret* and *secrete*). The new rule is to add only *-d*, unless the *e* in *-ed* is needed for correct pronunciation: *dated*, *pitted*, *added*, *learned* (JJ) and *dogged* (JJ).

Disambiguation of *gh*. The silent *gh* will be shortened to *h*. The preceding *o* or *u* will be dropped to maintain correct pronunciation. The *h* is retained to preserve vowel length as in *night* → *niht*, and to retain familiarity with the older forms. The *f*-sounding *gh* will be replaced by *ff* to preserve vowel lengths, e.g. *cough*, *coughing*, *coughed* will be *coff*, *coffing*, *coffd*. *Slough* will have three spellings: *slouh*, *sluh*, *shuff*. The latter two are already the Americanized spellings for *slough* meaning marsh/pond and shed/cast off, respectively. As shown later, *gh* will be used to indicate a hard *g*.

Disambiguation of *ea*. The string *ea* can represent four sounds. In case of a single-syllable *ea*, the sounds are **ɛ**, **i:**, and **ei**. There is no apparent rule for its pronunciation, and it is present in four homographs: *cleanly*, *read*, *lead*, and *tear*, each with two pronunciations. The new rule is that all *ea* forms will keep the **i:** sound, and other sounds will be represented by *e*, *eh*, or *ae*, as needed, e.g. *hedd*, *erth*, *behr*, *lether*, *graet*. Homographs will be split as needed, and in the case of *tears* there will be four spellings: *tears*, *tearz*, *taers*, *taerz*. For *bear*, the animal will be *baer* (same as the German spelling) and carry/tolerate will be *behr*, with a single-letter change.

Disambiguation of *o*. There are many commonly used words where *o* is used for the **ʊ** or **u:** sound: *to*, *do*, *lose*, *move*, *prove*, etc. These words will be modified on a case-by-case basis.

Disambiguation of *ou*. The *ou* form is present in about 9000 words in the Lexicon. These fall into three categories: native English words, adjective endings *-ous*, and French or Greek loan words. It can have four different pronunciations: **aʊ** (*out*), **u:** (*group*), **oʊ** (*soul*), and **ə** (*nimbus*). The schwa (**ə**) almost always occurs in the unstressed adjective ending *-ous*, where changing *-ous* to *-us* will create dozens of collisions (e.g. *callous*, *pious*, *venous*), so it remains unchanged. The **oʊ** sounds are derivatives of *soul*, and will be represented with *oh*. The **u:** sounding cases, the majority of which are in Greek and French loan words, will be mostly replaced by *oo* or *u* depending on the context, except for *oup* which is unchanged because the sound **aʊp** does not exist in English. *Four*, *tour*, and *pour* will be *forr*, *toor*, and *porr*. *Couth* and *youth* will become *cuth* and *yuth*, consistent with *Ruth* and *truth*. Words ending in *-outh* pronounced **-aʊθ** will remain unchanged.

The liberation of *q*. The restriction of *q* having to be followed by *u* will be lifted. In cases where the *u* in *qu* is silent, dropping it solves more than one problem: it allows the *q* to act as an unambiguous **k** sound, while preserving both familiarity and the length of the preceding vowel. Hence we will have *macaq*, *opaque*, *picturesq*, *techniqe*, *torq*, etc. Now all *qu* forms will have the **kw** sound. The other use of *q* is to use *qh* as the **tʃ** sound as in *church*, discussed below.

The doubling of *v*. The letter *v* seems to be treated differently from other consonants: first, there is a reluctance to end a word in *v*, resulting in vowel-length violations as in *active*, *give* and *have*. The spurious vowels in these words will be removed. Second, the rule of doubling a consonant to indicate a

preceding short vowel is not applied to *v*, except for a few more recent words like *chivvy*, *divvy*, *savvy*. A possible reason may have been to avoid confusing *vv* with *w*. This means that the vowel length before a *v* has to be memorized: *cover/over*, *seven/even*, *devil/evil*, *gavel/navel*, etc. Unfortunately, there are more than 500 words in this category, so the change has been restricted to *eav* → *evv*, only to disambiguate *ea*. Similar doublings in general may not be sound-neutral for some consonants, so *acid* will remain unchanged, as will *tether*, *cosher* and other words with two-letter consonants.

The new grapheme *qh*. The ambiguity of *ch* in English is most annoying, something which is absent in other languages. The Greek *chi* poses no problem for Spanish which writes it as *c/qu*, or other European languages which pronounce it as their native *ch*. In English the Greek *chi* is pronounced *k* with a few exceptions like *arch-* which can be **ark** or **artʃ**. In Euspell, words borrowed from Greek will retain the *ch* as the **k**-sound, and the **tʃ** sound will be represented by *qh*. The reason for this is to keep scientific words, where most *ch* forms represent the Greek *chi*, unchanged. Also, many other languages have their own *ch* pronunciation, so the English *qh* will be the unambiguous universal **tʃ** sound. The French *ch* will become *sh* as discussed later.

The new grapheme *uw*. There is currently no unambiguous way to represent the sound **au** as in *cow* in English. The form *uw*, inspired by Dutch *ouw*, is introduced to remedy this, replacing the ambiguous *ow*, but not the *ou* (as in *noun*). Instead, the **u:** sounding *ou*, almost entirely in French and Greek loan words, will be adjusted based on the context, as mentioned earlier. Some words will end up with two spellings, e.g. *bow/buw*, *sow/suw*. The Lexicon has four traditional words with the *uw* string: *juwansa*, *lauwine*, *rauwolfia*, and *vrouw*. In *juwansa*, *uw* is not a grapheme (the *w* belongs to the next syllable), and the pronunciation of *uw* in the last three actually matches the new form.

The soft *g*. Soft *g*'s are mostly of French origin, origin, also extended to Latin and Greek words. The rule is straightforward: *g* is soft before soft vowels *e*, *i*, and *y*, and hard before *a*, *o*, and *u*. Changing these to *j* will affect too many scientific terms – these are left alone. Words ending in *-inge* may drop the final *e* in their gerunds, spelled both as *binging/bingeing* or *hinging/hingeing*, but only *singeing*, because *singing* would be ambiguous. Words ending in *-geon* will lose the *o* and gain a *d* if needed, e.g. *bludgen*, *dundgen*, *pidgen*, because the pronunciations of *dge* and *dgi* are not ambiguous.

The hard *g*. Euspell maintains *gh* as hard *g*, in some cases introducing an *h* to enforce this, for example *rogue* → *roghe*, *guide* → *ghide*. This follows the unambiguous Italian convention (as does the now-consistently **k**-sounding *ch*) which we believe is more appropriate for English than the French/Spanish *gu*, because the *u* is pronounced in many words such as *distinguish*, *segue*, etc.

The silent *b*. The following common words in the Lexicon end in *mb* where the *b* is silent: *bomb*, *catacomb*, *climb*, *comb*, *crumb*, *dumb*, *jamb*, *lamb*, *limb*, *numb*, *plumb*, *recumb*, *rhumb*, *succumb*, *thumb*, *tomb*, and *womb*. In particular, *comb*, the only word with an **ou** sound, has no reasonable alternative spelling than *cohm*. Other words are changed based on the preceding vowel or whether the *b* is pronounced in derived words. So we have *clym*, *catacohm*, *cohm*, *dumm*, *jamm*, *lamm*, *limm*, *numm*, *thumm*, *toom*, and *woom*. Six words are unchanged because some of their derivatives have a non-silent *b*: *bomb*, *crumb*, *plumb*, *recumb*, *rhumb*, and *succumb*. The *b* in *debt* was inserted by Renaissance scribes imitating Latin *debitum*. *Debt* is reverted to *dett* as shown in Table 1.

The useful *h*. In some languages, a static *h* (one not followed by a vowel) is pronounced nevertheless, but not in English. This allows us to use this silent *h* to replace other silent letters while keeping

vowel lengths intact. *Fruit, suit, recruit, comb*, and *sew* will be *fruht, suht, recruht, cohm*, and *soh*. Many French words will have their silent *t*, *s*, and *x* replaced by *h*.

Words ending in -ost. The *o* in these words can be short as in *cost* and *lost*, or long as in *host* and *most*. The natural way of indicating a long *o* in these words is to add *e* to the end, as the old spellings of *most* and *ripost* are actually *moste* and *riposte*. Since *post* is also a prefix with over 700 words in the Lexicon – many of which are scientific terms – these are left alone, and only the root forms are changed, e.g. *poste*, *outposte*. All forms of *ghost* and *host* which are not already followed by a vowel will be changed to *goste* and *hoste*.

Doubling consonants. Some verb endings trigger the doubling of the last letter in British spelling, but not in American: *travelled/traveled*, *modelling/modeling*, etc. In case of the past tense, this is a moot point as both forms collapse into one: *traveld*, *canceld*, etc. For the VVG form, these will be made consistent using the following rule: if the final syllable is stressed, double the last letter before *-ing*: *compelling*, *referring* but *traveling*, *pivoting*. One caveat is that this rule applies to the base word, even if it receives a stressed prefix, e.g. *backfitting*, *typesetting*. Doubling letters in the stem was considered to indicate a preceding short vowel, affecting about 1400 words. Only a subset of these words are changed for other reasons, e.g. *heavy* → *hevvy* (to disambiguate *ea*), with others left alone to avoid inconsistency with derived forms, e.g. *study/studious*.

The *e* before -able. The suffix *-able* is currently added inconsistently to some words: *likable/likeable*, *sizable/sizeable*, etc. These will be made consistent with the following rule: if the *-ing* form drops the *e*, so should the *-able* form. This rule does not apply to words ending in *-ce* or *-ge*, to avoid hardening the consonants: *changeable*, *noticeable*, etc.

The final *th*. Words ending in *th/the* have a consistent rule in voicing: *-th* is voiceless and *-the* is voiced, e.g. *cloth/clothe*. The two exceptions are *betroth* and *smooth*, which will be *betrothe* and *smoothe*. The change in *betrothe* also fixes the long *o* problem.

French words. A large number of French words need to be respelled to conform with English. The strings *ch* and *oi* are pronounced **ʃ** and **wa** in French respectively, and will be spelled as *sh* and *wa*: *noir* → *nwar*, *machine* → *mashine*. The *gue* and *gui* forms will be *ghe* and *ghi* as mentioned earlier. *Boeuf* in particular has been left alone because it is pronounced three ways: **bu:f**, **boʊf**, and **buʃ**.

Foreign names. The general guideline for foreign names is to apply the Euspell rules only if they are transliterated from a different writing system. Latin-alphabet names are unchanged.

Singletons. Three words have a characteristic not present in any other English word. *I* is the only always-cap word in English, confused with the Roman numeral one and the letter *i* as intended. *I* will be normal-case like other words and spelled *ih*. Next is *of*, the only word with a voiced *f*. It will be spelled *ov*. The third is *shall*, the only word ending in *-all* but rhyming with *pal* (excluding unstressed syllables). It will be spelled *shal*.

Apps, extension, and add-ins

No spelling system can be of much use without aids to help in learning and usage. Euspell comes with a comprehensive set of software tools for reading and writing. Table 2 shows the platforms and software availability of Euspell. See Appendix A for installation and additional details.

Table 2. Platform availability

	Windows	MacOS	Linux	Android	iOS
Browser extension	Chrome, Edge, Firefox	Chrome, Safari, Firefox	Chrome, Firefox	–	–
PDF reader	in-browser, eupub	in-browser, eupub	in-browser, eupub	eupub	eupub
Text/epub reader	eupub	eupub (no Intel)	eupub	eupub	eupub
Office add-ins	MS Word, LibreOffice, Google Docs	MS Word, Google Docs, LibreOffice, Apple Pages (JXA)	LibreOffice, Google Docs	–	–
Dictation	Chrome, Edge	Chrome, Safari	Chrome	–	–
Game	Any browser				

 The Euspell logo is Unicode character 0x10F1, the Georgian letter ‘archaic he’ in blue (RGB 0000ff)

Browser extensions. The Euspell browser extensions convert web pages in real time, with the option of turning it on and off globally. Extensions are available for Chrome, Edge, Firefox, and Safari on desktops.

PDF. The PDF viewer is integrated into the browser extension on Windows, Mac and Linux. On the desktop and mobile devices, the eupub app also opens and converts PDFs.

Epub and text. A standalone app called eupub is created for reading text, epub, and PDF documents displayed in Euspell. It is available for Windows, Mac, Linux, Android, and iOS. PDF files are only shown converted.

Spelling checker. Spelling checker functionality is included in the word processor add-ins. A custom dictionary for new spellings and an exclusion dictionary are available.

Grammar checker. A grammar checker’s first step is POS tagging, and its hardest tagging cases are exactly the homographs Euspell reforms. The main Lexicon contains about 35000 words with new spellings. Many of these words can have more than one PoS, totaling about 46000, contained in the lexicon eupell_pos.tsv to incorporate into grammar checkers.

Speech synthesis. The eupell_tts.pls lexicon specifies the pronunciations of over 35000 words with new spellings in the IPA format, all unambiguous, allowing text-to-speech engines to pronounce new words. See more details at github.com/ossiak/euspell/blob/master/docs/ssml-lexicon.md .

Dictation. Lets a user speak and have their words inserted already spelled in Euspell — an authoring tool, the inverse of the reader. The recognizer produces traditional English; the existing Euspell converter respells it, one sentence at a time. For a detailed discussion and analysis of Euspell dictation, visit github.com/ossiak/euspell/blob/master/docs/dictation.md .

Game. A text-based game to practice your Euspell conversion. Instructions are shown on the web page.

How to support the project

The adoption of Euspell depends on a grassroots effort and support at multiple levels. Here are some ways to help the project:

- Install the software and get used to the new spelling – it is matter of days
- Share this paper and [#euspell](#) with others
- If you are a student, ask your instructors to accept your assignments in Euspell
- If you are the parent of a student, ask the school and the school district to recognize Euspell
- If you are an influencer, spread the word, and guide your followers to euspell.org
- If you are an NLP developer, incorporate Euspell in your apps.
- Lobby government agencies to adopt Euspell
- Wear the Euspell T-shirt and hoodie: shop.euspell.org.

Conclusion

The rules of English spelling are not hard to learn, were it not for the numerous exceptions. The main goal of Euspell is to minimize these exceptions and make spelling as consistent as possible within the current rules and spirit of English spelling, and to reduce ambiguities to make reading and comprehension easier with minimum change. The majority of new spellings are only one letter different from the original. A lexicon of around 205000 words, each carrying its parts of speech and its reformed spelling, already provides this consistency. Four fifths of the dictionary is untouched, no new letters or diacritics are introduced, and proper nouns are left alone. A classifier reading the surrounding words settles the entries that genuinely require context, resolving the dominant noun/verb ambiguity correctly in 94% of instances. Euspell was created in the hope of setting the groundwork for a grassroots effort to reform English spelling, without any destructive effect on existing documents. Billions of article, papers, and books are released electronically every year, and being able to read these through the Euspell filter allows everyone to judge it for themselves. This filter works in real time and can be switched on and off easily. Installation of the software is straightforward, and all lexicons, documentation, and code are open-sourced.

Future Work

A few issues are still unresolved. First is distinguishing the two short *u* sounds $\mathfrak{u}$ and $\wedge$ (*put* versus *putt*). The second case is the of spelling the hard and soft *g* sounds. Phonetic spelling of these, e.g. *get* → *ghet*, *gist* → *jist* is too disruptive, with hundreds of scientific words being affected. In both cases the phonetically correct answer costs more in legibility than it returns, and a partial fix would be worse than none. The only *gh* forms enforcing the hard *g* are those where *h* replaces *u* for the purpose of disambiguation. A third issue is statistical: the noun/verb classifier degrades on minimalist headlines and short phrases, where there is too little surrounding text to read. Beyond the reform itself, the machine-readable outputs are not explored yet. The pronunciation lexicon (about 35000 entries carrying IPA) and the part-of-speech lexicon (about 46000 entries in Penn Treebank tags) are both generated and awaiting integration into apps: the first exists so that a speech synthesizer can read converted text correctly, and the second is intended to teach grammar checkers like Harper [13] and

LanguageTool [14], a vocabulary they do not currently possess, rather than to justify building a checker of our own. Real-time spelling and grammar checking inside word processors remains blocked by the host applications. A fourth issue is vowel length consistency: hundreds of words could have been respelled to indicate short vowels which currently may be seen as long, e.g. *seven* → *seven*, *honor* → *honnor*, *epic* → *eppic*. This change would be too disruptive to implement at this time.

AI assistance. The reform architecture and the main lexicon are created by the author. Anthropic Claude [15] Opus 4.8/5 and Fable 5 assisted with software development, revisions to sections of this paper, installation instructions, website design, and video production. Machine learning is used only to decide which of two or more spellings, when applicable, is chosen in a given context.

Trademark. EUSPELL is a trademark of Kamran Ossia; United States registration is pending. The mark identifies this project and its software. Using the word for the spelling itself — describing text as being in euspell, or writing about the reform — needs no permission and is not restricted.

References

- [1] University of Lancaster PoS tagset: ucrel.lancs.ac.uk/claws7tags.html
- [2] Penn Treebank: www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- [3] International Phonetic Alphabet: www.internationalphoneticassociation.org/content/ipa-chart
- [4] Mulcaster, R. “The First Part of the Elementarie”, 1582.
ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/3176
- [5] George Bernard Shaw’s Shavian alphabet, en.wikipedia.org/wiki/Shavian_alphabet
- [6] Webster, N. “A Compendious Dictionary of the English Language”, 1806.
archive.org/details/compendiousdictionaryoftheenglishlanguage1806
- [7] Paine, H.G. “Handbook of Simplified Spelling”, 1920.
www.spellingsociety.org/files/handbook-of-simplified-spelling-1546879176.pdf
- [8] Upward, C. “Cut Spelling: A Handbook”, 1996. www.spellingsociety.org/uploaded_books/cs1intro.pdf
- [9] Seymour, P.H.K., Aro, M., Erskine, J.M. “Foundation literacy acquisition in European orthographies”, *British Journal of Psychology* 94, 2003, pp. 143–174.
- [10] Landerl, K., Wimmer, H., Frith, U. “The impact of orthographic consistency on dyslexia: A German-English comparison”, *Cognition* 63, 1997, pp. 315–334.
- [11] Paulesu, E. *et al.* “Dyslexia: Cultural diversity and biological unity”, *Science* 291, 2001, pp. 2165–2167.
- [12] KPMG Foundation. “The long term costs of literacy difficulties”, December 2006.
- [13] Harper grammar checker: writewithharper.com
- [14] LanguageTool: languagetool.org
- [15] Claude: claude.ai

Appendix A. Installing the tools

Everything described in this paper is free software under GPL-3.0 or later for code and Creative Commons Attribution-ShareAlike 4.0 for the lexicons. Everything runs on the reader’s own device. There is no account, no server, and no telemetry: each tool carries its own copy of the lexicon, and every conversion is performed locally.

A.1 Downloads

euspell.org carries the current download links for the tools which have reached a store or a release. The browser extension is on the Chrome web store, also covering Edge, Brave, Opera and other Chromium browsers: chromewebstore.google.com/detail/euspell/jjibndkdmbmomfmngblomkkejgdnemja .

The software and its documentation are in two public repositories:

Repository	What it holds
github.com/ossiak/euspell	The lexicon, the conversion engine, the browser extension, and the word-processor add-ins
github.com/ossiak/eupub	Eupub, the standalone EPUB, PDF and plain-text reader

A.2 Instructions

Each tool has one install document, kept current with the authoritative version of the code it describes. Where they disagree with this paper, they prevail.

To install	Read
The browser extension — Chrome, Edge, Brave, Opera, Vivaldi, Firefox, Safari	installing.md in github.com/ossiak/euspell/tree/master/docs
The word-processor add-ins — Microsoft Word, LibreOffice Writer, Google Docs, Apple Pages	installing-addins.md in github.com/ossiak/euspell/tree/master/docs
Eupub , the e-reader — Windows, macOS, Linux, Android, iOS	installing.md in github.com/ossiak/eupub/tree/main/docs
The conversion game — any modern browser	Nothing to install: visit euspell.org/game/

All the builds themselves are available on a single page euspell.org/downloads on the website. For the browser extension: Firefox xpi, the notarized Safari dmg, and a link to the Chrome Web Store listing. For Eupub, a signed Windows installer, a notarized macOS dmg, a Linux AppImage, a signed Android APK, and a link to the App Store listing for iPhone. The same files are available from the github repository.

A.3 Current status

As of August 2026, here is the status of the tool set:

- **The browser extension** is on the Chrome Web Store, which also serves Edge, Brave, Opera and other Chromium browsers. Safari has no store listing but does have a signed, notarized app to download, drag to Applications, and switch on in Safari's settings. Firefox has neither a listing nor a build step: its Mozilla-signed .xpi is published with the releases, and opening that file in Firefox installs it permanently.
- **The word-processor add-ins** have no marketplace listing on any of the four platforms. Three are built from source; Word can instead be pointed at a hosted manifest and needs no build at all. They are one-pass converters rather than live spell-checkers: a command rewrites the document, and nothing is underlined as you type.
- **Eupub** ships prebuilt for the three desktop platforms: Windows (.exe), macOS (.dmg), and Linux (.AppImage). For Android, the .apk is signed with the release key and installs directly. For iOS 17 or later, search the App Store.
- **The conversion game** is the one item with nothing to install: a single web page that asks the reader to rewrite a passage in Euspell by hand and scores what they produce. It needs no server, no build, and no network, and it is the quickest way to find out how easy the reform is to learn.

A.4 If a link has moved

The repositories are the durable items, and each is self-sufficient: the instructions live in the **docs/** directory of the repository they describe, so a clone or a downloaded ZIP carries the software and its documentation together. A reader who can reach either repository needs nothing else from this appendix, and one who can reach neither will find the project's current address through a search for *euspell*.

Appendix B. CLAWS7 [1] and Penn TreeBank [2] part-of-speech tagsets

CLAWS7	PTB	Description
APPGE	PRP\$	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	DT	article (e.g. the, no)
AT1	DT	singular article (e.g. a, an, every)
BCL	RB	before-clause marker (e.g. in order (that), in order (to))
CC	CC	coordinating conjunction (e.g. and, or)
CCB	CC	adversative coordinating conjunction (but)
CS	IN	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	IN	as (as conjunction)
CSN	IN	than (as conjunction)
CST	IN	that (as conjunction)
CSW	IN	whether (as conjunction)
DA	DT	after-determiner or post-determiner capable of pronominal function
DA1	DT	singular after-determiner (e.g. little, much)
DA2	DT	plural after-determiner (e.g. few, several, many)
DAR	JJR	comparative after-determiner (e.g. more, less, fewer)
DAT	JJS	superlative after-determiner (e.g. most, least, fewest)
DB	PDT	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	PDT	plural before-determiner (both)
DD	DT	determiner (capable of pronominal function) (e.g. any, some)
DD1	DT	singular determiner (e.g. this, that, another)
DD2	DT	plural determiner (these, those)
DDQ	WDT	wh-determiner (which, what)
DDQGE	WP\$	wh-determiner, genitive (whose)
DDQV	WDT	wh-ever determiner, (whichever, whatever)
EX	EX	existential there
FO	SYM	formula
FU	SYM	unclassified word
FW	FW	foreign word
GE	POS	germanic genitive marker - (' or 's)
IF	IN	for (as preposition)
II	IN	general preposition
IO	IN	of (as preposition)
IW	IN	with, without (as prepositions)
JJ	JJ	general adjective
JJR	JJR	general comparative adjective (e.g. older, better, stronger)
JJT	JJS	general superlative adjective (e.g. oldest, best, strongest)
JK	JJ	catenative adjective (able in be able to, willing in be willing to)
MC	CD	cardinal number, neutral for number (two, three..)
MC1	CD	singular cardinal number (one)
MC2	CD	plural cardinal number (e.g. sixes, sevens)
MCGE	CD	genitive cardinal number, neutral for number (two's, 100's)
MCMC	CD	hyphenated number (40-50, 1770-1827)
MD	JJ	ordinal number (e.g. first, second, next, last)
MF	CD	fraction, neutral for number (e.g. quarters, two-thirds)
ND1	NN	singular noun of direction (e.g. north, southeast)

NN	NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	NN	singular common noun (e.g. book, girl)
NN2	NNS	plural common noun (e.g. books, girls)
NNA	NN	following noun of title (e.g. M.A.)
NNB	NN	preceding noun of title (e.g. Mr., Prof.)
NNL1	NN	singular locative noun (e.g. Island, Street)
NNL2	NNS	plural locative noun (e.g. Islands, Streets)
NNO	CD	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	NNS	numeral noun, plural (e.g. hundreds, thousands)
NNT1	NN	temporal noun, singular (e.g. day, week, year)
NNT2	NNS	temporal noun, plural (e.g. days, weeks, years)
NNU	NN	unit of measurement, neutral for number (e.g. in, cc)
NNU1	NN	singular unit of measurement (e.g. inch, centimetre)
NNU2	NNS	plural unit of measurement (e.g. ins., feet)
NP	NNP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	NNP	singular proper noun (e.g. London, Jane, Frederick)
NP2	NNPS	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	NNP	singular weekday noun (e.g. Sunday)
NPD2	NNPS	plural weekday noun (e.g. Sundays)
NPM1	NNP	singular month noun (e.g. October)
NPM2	NNPS	plural month noun (e.g. Octobers)
PN	NN	indefinite pronoun, neutral for number (none)
PN1	PRP	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)
PNQO	WP	objective wh-pronoun (whom)
PNQS	WP	subjective wh-pronoun (who)
PNQV	WP	wh-ever pronoun (whoever)
PNX1	PRP	reflexive indefinite pronoun (oneself)
PPGE	PRP\$	nominal possessive personal pronoun (e.g. mine, yours)
PPH1	PRP	3rd person sing. neuter personal pronoun (it)
PPHO1	PRP	3rd person sing. objective personal pronoun (him, her)
PPHO2	PRP	3rd person plural objective personal pronoun (them)
PPHS1	PRP	3rd person sing. subjective personal pronoun (he, she)
PPHS2	PRP	3rd person plural subjective personal pronoun (they)
PPIO1	PRP	1st person sing. objective personal pronoun (me)
PPIO2	PRP	1st person plural objective personal pronoun (us)
PPIS1	PRP	1st person sing. subjective personal pronoun (I)
PPIS2	PRP	1st person plural subjective personal pronoun (we)
PPX1	PRP	singular reflexive personal pronoun (e.g. yourself, itself)
PPX2	PRP	plural reflexive personal pronoun (e.g. yourselves, themselves)
PPY	PRP	2nd person personal pronoun (you)
RA	RB	adverb, after nominal head (e.g. else, galore)
REX	RB	adverb introducing appositional constructions (namely, e.g.)
RG	RB	degree adverb (very, so, too)
RGQ	WRB	wh- degree adverb (how)
RGQV	WRB	wh-ever degree adverb (however)
RGR	RBR	comparative degree adverb (more, less)
RGT	RBS	superlative degree adverb (most, least)
RL	RB	locative adverb (e.g. alongside, forward)

RP	RP	prep. adverb, particle (e.g. about, in)
RPK	IN	prep. adv., catenative (about in be about to)
RR	RB	general adverb
RRQ	WRB	wh- general adverb (where, when, why, how)
RRQV	WRB	wh-ever general adverb (wherever, whenever)
RRR	RBR	comparative general adverb (e.g. better, longer)
RRT	RBS	superlative general adverb (e.g. best, longest)
RT	RB	quasi-nominal adverb of time (e.g. now, tomorrow)
TO	TO	infinitive marker (to)
UH	UH	interjection (e.g. oh, yes, um)
VB0	VBP	be, base form (finite i.e. imperative, subjunctive)
VBDR	VBD	were
VBDZ	VBD	was
VBG	VBG	being
VBI	VB	be, infinitive (To be or not... It will be ..)
VBM	VBP	am
VBN	VBN	been
VBR	VBP	are
VBZ	VBZ	is
VD0	VBP	do, base form (finite)
VDD	VBD	did
VDG	VBG	doing
VDI	VB	do, infinitive (I may do... To do...)
VDN	VBN	done
VDZ	VBZ	does (verb, not multiple female deer)
VH0	VBP	have, base form (finite)
VHD	VBD	had (past tense)
VHG	VBG	having
VHI	VB	have, infinitive
VHN	VBN	had (past participle)
VHZ	VBZ	has
VM	MD	modal auxiliary (can, will, would, etc.)
VMK	MD	modal catenative (ought, used)
VMXX	VM RB	can't, cannot
VV0	VBP	base form of lexical verb (e.g. give, work)
VVD	VBD	past tense of lexical verb (e.g. gave, worked)
VVG	VBG	ing participle of lexical verb (e.g. giving, working)
VVGK	VBG	ing participle catenative (going in be going to)
VVI	VB	infinitive (e.g. to give... It will work...)
VVN	VBN	past participle of lexical verb (e.g. given, worked)
VVNK	VBN	past participle catenative (e.g. bound in be bound to)
VVZ	VBZ	s form of lexical verb (e.g. gives, works)
XX	RB	not, n't
ZZ1	NN	singular letter of the alphabet (e.g. A,b)
ZZ2	NNS	plural letter of the alphabet (e.g. A's, b's)

Appendix C. International Phonetic Alphabet (IPA) for English [3]

ɑ	father
ɒ	lot
æ	tap
aɪ	nice
aʊ	cow
ɛ	leg
ɜ	bird
eɪ	take
ɪ	big
i	only
i:	meet
oʊ	goat
ɔ	thought
ɔɪ	boy
ʊ	foot
u:	goose
ʌ	but
ə	about

b	boy	g	go	l	low	r, ɹ	rare	θ	thing
d	dad	h	head	m	may	s	so	v	vive
dʒ	joy	hw	where	n	nun	ʃ	shall	w	wine
ð	they	j	yes	ŋ	sing	t	tote	z	zoo
f	fall	k	kick	p	pop	tʃ	church	ʒ	measure

Stress markers applied before syllables

Primary stress	' ____
Secondary stress	ˈ ____

Appendix D. Euspell encodings and their distribution in the Lexicon

Encoding	Description	Count
000	unchanged common word	158822
011	unambiguous VVZ ending $-s \rightarrow -z$	5981
012	NN2 VVZ $-s$ ending disambiguated	4916
021	unambiguous JJ VVD VVN ending $-ed \rightarrow -d$ when warranted	4699
022	JJ VVD VVN ending disambiguated e.g. <i>blessd/blessed</i>	5
041	doubling consonant before ending	3
101	stem change	14589
102	stem disambiguated, (most ending in <i>-ate</i>)	155
103	stem disambiguated three ways	5
111	stem change + unambiguous VVZ ending $-s \rightarrow -z$	1462
112	stem change + NN2 VVZ ending – disambiguated	746
113	stem change + NN2 VVZ ending – disambiguated three ways	4
114	stem change + NN2 VVZ ending – disambiguated four ways	5
121	stem change + JJ VVD VVN ending $-ed \rightarrow -d$ when warranted	953
123	stem change + JJ VVD VVN ending – disambiguated three ways	1
131	stem change, undoubling consonant before ending	21
152	stem disambiguated, not ending in <i>-ate</i>	15
202	stem change – semantic disambiguation needed	59
500	unchanged rare or archaic word	5087
501	stem change – rare or archaic word	85
511	VVZ ending $-s \rightarrow -z$ – rare or archaic word	295
521	JJ VVD VVN ending $-ed \rightarrow -d$ when warranted – rare or archaic word	131
601	merges with existing word (mostly British $\rightarrow$ American)	5980
631	merges with existing word – undoubling consonant before ending	307
641	merges with existing word – doubling consonant before ending	42
700	unchanged word – French pronunciation	115
701	stem change – French pronunciation	892
702	NN1 NN2 disambiguation – French pronunciation	9
711	VVZ ending $-s \rightarrow -z$ – French pronunciation	4
721	JJ VVD VVN ending $-ed \rightarrow -d$ when warranted – French pronunciation	26
800	unchanged word – Scottish pronunciation	43
811	VVZ ending ending $-s \rightarrow -z$ – Scottish pronunciation	2
821	VVD VVN ending $-ed \rightarrow -d$ when warranted – Scottish pronunciation	6
900	abbreviation	40