

# Introduction to the CommonWords Database

## October, 2010

### Fields in the CommonWords Table:

- Word
- Correspondences: Sound-to-Spelling and Spelling-to-Sound
  - <Vre> Spellings
  - Syllabic Consonants
  - An Update on Low Back Vowels
- Explication
- Analysis
- Themes
- Homophones
- Homographs
- Other Problem Spellings
- Spelling Difficulty
- Rank
- Range
- Subrange
- Characters
- Syllables
- Syllable Structure
- Complexity
- Prefixes
- Suffixes
- Parts of Speech
- Sources

### Other Data Tables:

- Correspondences: Sound to Spelling
- Correspondences: Spelling to Sound
- Sounds Count
- Letters Count
- Spellings Count

\* \* \* \* \*

The CommonWords database consists of seven data tables: (i) CommonWords, a list of 8591 words, at least 6500+ of which are high-frequency; (ii) Correspondences: Sound to Spelling, a list of the 356 sound-to-spelling correspondences found in those 8591 words; (iii) Correspondences: Spelling to Sound, a list of the 356 spelling to sound correspondences, (iv) Themes, a list of 169 themes, or topics, for which more than 7200 words in the CommonWords table have been tagged, (v) Sounds Count, a list of the frequency of occurrence of the 68 sounds in the 8591 words in CommonWords, (vi) Letters Count, a list of the frequency of occurrence of the 26 letters, and (vii) Spellings Count, a list of the frequency of the 155 different spellings recognized in CommonWords. More details on each of these seven tables are provided below.

**CommonWords.** This, the main table in the database, contains the following fields:

**Word.** The Word field can be used to filter to words with various letter strings – for instance, “Word contains sh” returns all words with the consonant digraph <sh> anywhere in the word (plus *grasshopper* , grass+hop+p+er]01, in which the <sh> is not a digraph). “Word ends with sh” returns only those words with final <sh>.

You can filter, among other things, for four different kinds of consonant strings:

1. The following fifty consonant blends – that is, strings of two or three consonant letters that spell two or more consonant sounds in the same syllable. Some blends are word-initial, more are word-final, a few are both. For instance, “Word contains rm” returns 93 words, some of which are false hits, the <r> and <m> being divided by a syllable or element boundary as in *chairman* . However, all of the blends listed below are tagged “cb” (consonant blend) in the Analysis column. To avoid false hits, you can filter to, say, “Word contains bl” and “Analysis contains cb”, thus avoiding the many words ending in <ble> in which a schwa sound occurs between the [b] and the [l], as in *able*:

<bl> = [bl] as in *blue*  
<br> = [br] as in *brew*  
<chr> = [kr] as in *chronicle*  
<cl> = [kl] as in *clue*  
<cr> = [kr] as in *crew*  
<ct> = [kt] as in *act*  
<dr> = [dr] as in *draw*  
<fl> = [fl] as in *flaw*  
<fr> = [fr] as in *from*  
<ft> = [ft] as in *soft*  
<gl> = [gl] as in *gloom*  
<gr> = [gr] as in *groom*  
<ld> = [ld] as in *sold*  
<lf> = [lf] as in *shelf*  
<lt> = [lt] as in *belt*  
<mp> = [mp] as in *camp*  
<nch> = [nch] as in *branch*  
<nd> = [nd] as in *brand*  
<nk> = [ŋk] as in *sank*  
<nt> = [nt] as in *sent*  
<nth> = [nth] as in *tenth*  
<pl> = [pl] as in *place*  
<pr> = [pr] as in *price*  
<pt> = [pt] as in *slept*  
<qu> = [kw] as in *quarter*  
<rch> = [rch] as in *march*

<rd> = [rd] as in *hard*  
<rk> = [rk] as in *mark*  
<rl> = [rl] as in *girl*  
<rm> = [rm] as in *arm*  
<rn> = [rn] as in *barn*  
<rt> = [rt] as in *short*  
<rth> = [rth] as in *birth*  
<sc> = [sk] as in *scale*  
<sch> = [sk] as in *school*  
<scr> = [skr] as in *scrape*  
<shr> = [shr] as in *shrill*  
<sk> = [sk] as in *ask*  
<sl> = [sl] as in *sled*  
<sm> = [sm] as in *small*  
<sn> = [sn] as in *sneak*  
<sp> = [sp] as in *spoke*  
<sph> = [sf] as in *sphere*  
<spl> = [spl] as in *splash*  
<spr> = [spr] as in *spring*  
<squ> = [skw] as in *squeak*  
<st> = [st] as in *last*  
<str> = [str] as in *straw*  
<sw> = [sw] as in *swell*  
<thr> = [thr] as in *throw*  
<tr> = [tr] as in *true*  
<tw> = [tw] as in *twin*  
<tz> = [ts] as in *quartz*

2. You can also filter to the following consonant doublets and doublet equivalents, which spell a single consonant sound, usually after a short vowel, nearly always in word-medial position, and usually due to the twinning of a final consonant when adding a suffix (as in *twinning*) or the assimilation of final consonants in prefixes (as in *announce*):

<bb> = [b] as in *robber*  
<cc> = [k] as in *accurate*  
<ck> = [k] as in *rock*  
<cq> = [k] as in *acquire*  
<dd> = [d] as in *reddest* or *odd*  
<dg> = [j] as in *bridge*  
<dj> = [j] as in *adjourn*  
<ff> = [f] as in *offer*  
<gg> = [g] as in *bigger* and *egg*  
<ll> = [l] as in *allow* and *tell*  
<mm> = [m] as in *hammer*

<nn> = [n] as in *dinner*  
<pp> = [p] as in *happy*  
<rr> = [r] as in *carry*  
<ss> = [s] as in *missing*  
<tch> = [ch] as in *catch*  
<tt> = [t] as in *cotton*  
<zz> = [z] as in *dizzy*

3. You can filter to consonant digraphs or trigraphs – two or three consonant letters that spell a single consonant sound – including doublets like <bb>. Others:

<ch> = [ʃ] as in *church* and *echo*  
<gh> = [f] as in *laugh* or [g] as in *ghost*  
<ght> = [t] as in *night*  
<ph> = [f] as in *phone*  
<rh> = [r] as in *rhyme*  
<rrh> = [r] as in *myrrh*  
<sh> = [ʃ] as in *shirt*  
<tch> = [tʃ] as in *witch*  
<th> = [θ] and [ð] as in *thin* and *this*  
<wh> = [h] as in *whole*  
<wh> = [w] (or [ʰw]) as in *while*  
<wr> = [r] as in *write*

4. And you can filter to the following simplifications, which retain the original longer spellings of one-time blends that have simplified over time to single consonant sounds:

<cht> = [t] as in *yacht*  
<ft> = [f] as in *often*  
<ght> = [t] as in *light*  
<gn> = [n] as in *sign*  
<kn> = [n] as in *knight*  
<ld> = [d] as in *could*  
<lf> = [f] as in *half*  
<lk> = [k] as in *talk*  
<lm> = [m] as in *calm*  
<ln> = [n] as in *Lincoln*  
<mb> = [m] as in *bomb*  
<mn> = [m] as in *column*  
<pb> = [b] as in *cupboard*  
<ph> = [p] as in *shepherd*  
<ps> = [s] as in *psychology*  
<qu> = [k] as in *conquer*  
<sc> = [s] as in *muscle*  
<sl> = [l] as in *island*  
<st> = [s] as in *listen*

<sth> = [s] as in *isthmus*  
<sw> = [s] as in *sword*  
<tg> = [g] as in *mortgage*

**Correspondences.** The Correspondences: Sound to Spelling field gives all sound to spelling correspondences found in each word, in order. The Correspondences: Spelling to Sound field gives all spelling-to-sound correspondences. The Correspondences: Sound to Spelling field is primarily for teachers of spelling and writing; the Correspondences: Spelling to Sound field is primarily for teachers of reading.

In both the Sound to Spelling and the Spelling to Sound fields square brackets enclose sounds, arrowhead brackets enclose spellings, and the equal sign translates to “is spelled” or “spells”. Thus, in the Sound to Spelling field “[k]=<c>” translates to “the sound [k] is spelled with the letter <c>”, and in the Spelling to Sound field “<c>=[k]” translates to “the letter <c> spells the sound [k]”. Curly braces mark silent letters: {D} marks silent letters that serve some diacritical function; {ND} marks silent letters with no diacritical function. Thus “{D}=<e>” indicates a diacritical silent <e>, as in *time, clothe, ounce, bronze, clause, league, active*, while {ND}=<e> indicates a non-diacritical silent <e>, as in *fixed*. and with the final <e> in *feature*.

The Sound to Spelling and Spelling to Sound fields in the CommonWords table allow various kinds of searches and filters. If you are dealing with phonics, you can filter to certain correspondences or to certain sounds or to certain spellings. For instance, at *meadow* the Sound-to-Spelling field contains the following: “[m]=<m> [e1]=<ea> [d]=<d> [o2]=<ow>”. Thus, you can filter to all the words in which short <e> is spelled <ea> (“[e1]=<ea>, 73 words”), or to all words that contain short <e> (“[e1]”, 1363), however it's spelled, or to the <ea> spelling (“<ea>”) – sometimes spelling short <e> (as in *meadow*), sometimes long <e> (as in *streak*), sometimes long <a> (as in *steak*), and sometimes schwa (as in *ocean*) (252 words).

**<Vre> Spellings.** The <Vre> spellings introduce some complexities. Because of the effect of the [r] on the preceding vowel, we must extend the notion of diacritic function to include sometimes subtle changes in the vowel sound other than the typical short-long distinction. In the <are> and <ire> spellings, the silent <e> is always diacritic: The <are> spelling always spells the [a3r] sound, as in *bare*. Since words like *bare* contrast in pronunciation with words like *bar*, the silent <e>'s can be said to be diacritic: *care/car, fare/far, mare/mar, pare/par, tare/tar*, etc. A similar pattern holds for the <ire> spelling, which consistently spells the [i2r] sound, as in *fire*. *Fire* contrasts with *fir* – as does *sire* with *sir*, so again the <e> is diacritic.

The opposite holds with the <ore> spelling, which consistently spells [o5r], as in *tore*. But phonetically *tore* does not contrast with the homophonic *tor* “rocky peak”, without the final <e>. or *torch* with a consonant after the <r>. So the <e> in <ore> is not diacritic. Other examples are *bore, border; for, fore; or, ore; sore, sort*, etc.

With the <ere> and <ure> spellings whether the silent final <e> is diacritic depends on

the sound being spelled. When <ere> spells [u1r] as in *were*, the final <e> is not diacritic: It does not affect the vowel sound, as can be seen by comparing words like *her* and *term*. However, when <ere> spells [e3r] as in *here*, the final <e> is diacritic – compare *here* and *her*. And when it spells [a3r] as in *there* and *where* and in the preposition *ere*, the final <e> is also diacritic. *Ere* contrasts with the interjection *er*, and *there* contrasts with, say, *thermostat*.

Similarly, when <ure> spells the unstressed [u4r], as in *capture*, it is not diacritic, as can be seen by comparing it with the word *sulphur* without the final <e>. However, when <ure> spells [u3r], as in *sure*, the <e> is diacritic – compare the word *surly* with no <e>. And when it spells [yu3r], as in *cure*, the <e> is also diacritic – compare *cure* with *cur*. There is room in any description of correspondences for honest differences of opinion, especially in view of the sometimes large differences in pronunciation among various dialects. These differences might be expected to arise particularly with the treatment of the non-long <o> vowels, schwa, and [r]-colored vowels.

**Syllabic Consonants.** Though the peak of most syllables is a vowel sound spelled by a vowel letter, in some syllables the peak is spelled by a consonant. These peak consonants are called *syllabic consonants*. The most common is [l2], usually spelled <l> as in *bottle*, but occasionally <ll> as in *satellite*. The nasal [n2] also can be syllabic, as in *button*. The nasal [m] is technically not syllabic since it always is preceded by a distinct schwa sound, as in *chasm* and *capitalism*. In some pronunciations syllabic consonants will be preceded by a full schwa, sometimes only a trace, sometimes none at all. And there tends to be a great deal of variation so far as the schwa sound is concerned, variation from dialect to dialect, from speaker to speaker within dialects, and even variation in a given speaker's pronunciation depending on the total context and situation. In CommonWords the tags [l2] and [n2] are used regardless of the degree of schwa sound possible in the syllable. "Syllabic" [m] is always represented as [u4m].

**An Update on Low Back Vowels.** In the earlier version of CommonWords, as in *American English Spelling (AES)*, I recognized only two different low back vowels, in *AES* called high and low short <o> and in the earlier CommonWords tagged as [o1] and [o4]. Doing so conflated some vowel sounds and spellings that I now believe should be distinguished. This new version of CommonWords recognizes three low back vowels, tagged [o1], [o3], and [o4]. (Long <o> continues to be tagged [o2].) I've used the *American Heritage Dictionary (AHD)* and *Webster's 3<sup>rd</sup> (W3)* as references since they are widely available and accessible. In *AHD*'s pronunciation system [o1] = [ɔ̄], [o3] = [ä], and [o4] = [ô]. In *W3*'s pronunciation system [o1] and [o3] both = [ä], and [o4] = [ô].

Distinguishing the three groups of low back vowels raises some difficulties, especially in identifying orthographically short and long vowels. Most long vowels are phonetically tense – that is, pronounced with marked tenseness in the vocal tract. Orthographically short vowels are phonetically lax.

The distinction between long and short vowels involves us with variations in pronunciation that go back into Middle English. For instance, in this revised

CommonWords there are 119 words with vowels tagged [o4] and spelled with the digraphs <au>, <aw>, <oa>, or <ou>. Since the orthographic tactical patterns like VCV and VCC do not apply to vowel digraphs, these 119 instances of [o4] lie outside the normal short and long tactics. I treat [o4] as linguistically tense but orthographically neither long nor short. The several instances with tense [o4] pronunciations listed first in the *AHD* and *W3* have variant pronunciations with short and lax [o1]. Those instances in which the vowel is spelled <o> I have tagged [o1] because the <o> spellings tend to be consistent with more general short vowel tactics. For instance, they typically occur in VCC or VC# patterns – as in *coffee* and *fog*. Also they are affected by silent final <e>'s as in the contrasts between *dog* and *doge*, *cloth* and *clothe*, *lop* and *lope*. Such is not the case with the <a> and digraph spellings, which I tag as [o4]. The degree of variation here is also reflected in the fact that seventeen other words with [o1] pronunciations listed first in the *AHD* have [o4] as variants: nine with the <o> spelling, eight with <a> following <w>. All of these I tag as [o1].

The low back vowel [o1] is the regular short <o>: In the CommonWords sample it is overwhelmingly spelled <o>: in 732 of the 771 instances of [o1]. It is spelled <a> in 32 words, always following [w], a regular rounding of the earlier <a> vowel that became widespread by the 17<sup>th</sup> century. In four words it is spelled <oh> or <ow>, as in *John* and *knowledge*. (The other four spellings of [o1] are in the adoptions *bureaucracy*, *entrepreneur*, *entourage*, and *leprechaun*.) In general the [o1] sounds occur in normal short vowel patterns VCC and CVC# and participate in normal short vs. long vowel contrasts, as in *wan* vs. *wane*, *dot* vs. *dote*, and *hop* vs. *hope*.

Orthographically, the low back vowels [o3], and [o4], are in a class quite different from [o1]. Both are tense, or at least moderately tense, vowels. And tense vowels are typically classed as orthographically long. But [o3] and [o4] do not behave like long vowels. Admittedly, they can occur in word-final position in words like *grandma* and *Panama*, *straw* and *jaw*, like other long (tense) vowels. But they both also occur in what would normally be short vowel patterns, as [o3] in *calm*, *father*, and [o4] in *salt* and *strong* (compare *film*, *rather*, *silt*, *string*). Neither [o3] nor [o4] seems ever to be involved in the normal tactics for long vowels. Thus, I treat [o3] and [o4] as tense, but not orthographically long. So we have lax and short [o1], tense and long [o2], and tense but neither long nor short [o3] and [o4].

For now, following Ladefoged in his *A Course in Phonetics* (p. 74), I treat the diphthongs [oi] and [ou] as tense, though for our purposes the issue is not too important since digraph spellings are not restrained by the standard tactical distinctions.

Though there are several complications, I regularly treat the consonant digraphs <ch>, <ph>, <sh>, and <th> as cases of CC in VCC strings. (See *AES*, 101-108 for discussion of some of the complications.)

**Explication.** Explication is the analysis of written words into their elements, or

smallest meaning-bearing parts. Thus explication analyzes words into their prefixes, bases, and suffixes. It also shows any deletions, insertions, or replacements that occur when the elements combine – for instance, final <e> deletion in *hast/e+y*] at *hasty*; final consonant insertion in *twin+n+ing*] at *twinning*, and replacement in [*a/d+p+pear*] at *appear*. In nearly all cases this field contains the same explications as given in the Words table of the larger Lexis database elsewhere on this site. The occasional differences are due to the different intended audiences for the two tables. Lexis is more for scholars and linguists; CommonWords is more for teachers and students. In Lexis the major desideratum is economy; in CommonWords it is accessibility, so a CommonWords explication will sometimes explicate to a free word where Lexis explicates to a more obscure, but economical, bound base. For more information on elements given in Explication, you should consult the appropriate tables in the Lexis database.

This Explication field can be used to filter to words with various prefixes, bases, suffixes, and procedures. The following are some possible search strings:

To find words that contain the prefix *de-*: "[de+"

To find words that contain the base *fect*: "fect"

To find words that contain the verbal suffix *-ing*: "+ing]1"

To find words that contain final <e> deletion: "/e+[eiouy]"

To find words that contain prefixes with assimilation: "/[!aeiouwy]+[!aeiouwy]+"

If you want to find words with given letter strings within a single element, you can filter for them in the Explication field: Thus, "Explication contains sh" returns only words that contain <sh> in a single element and does not return cases of <sh> due to concatenation, as in *grasshopper*. For more on explication see the article "Explication, Evolution, and Orthography" in the Short Articles section of this website – especially the final two sections, "Post-Alphabetic Orthography" and "Problems in Explication".

The **Analysis** field contains a number of the orthographically significant features of each word, each of which can be filtered to. For more details see the reference to *American English Spelling (AES)* given below in parentheses:

**y>i** = Instance of <y> to <i> change, as in *tries* from *try*. Also includes words with derived forms that would involve <y> to <i> changes and instances of <y> deletion (*AES*, 84-87)

**i>y** = Instance of <i> to <y> change, as in *lying* from *lie*. Also includes instances of <i> deletion (*AES*, 157)

**SWR** = Instance of Short Word Rule, as in *egg* and *pie*, with double final consonant or silent final <e> added to avoid words of less than three letters (*AES*, 87-89)

**VCC** = Stressed short vowel+consonant+consonant, as in *lettuce* (*AES*, 96-107). When not syllable-initial <x> is treated here as two consonants; thus *tax* and *taxi* are tagged as containing a VCC string.

**VCCX** = VCC holdout, as in *blind* (AES, 101-11)

**VCV** = Stressed long vowel+consonant+vowel (including silent final <e>), as in *vapor* and *rate* (AES, 96-100, 107-11)

**VCVX** = VCV holdout, as *done* (AES, 107-11)

**VrV** and **Vrr** are versions of the VCV and VCC patterns that involve the consonants <r> and [r]. The normal VCV/VCC contrast, as in *fate* and *fatten*, holds in several cases when the consonant involved is [r]. (For more details see “<Vre> Spellings” above.) However, we must allow for varying effects and degrees of effect of the [r] on the preceding vowel and must sometimes choose among accepted variant pronunciations. Such choosing is consistent with the Principle of Preferred Regularity: When given a choice of pronunciations (or spellings), we should choose the one that fits the pattern, that is most regular (AES, 25-26):

<arV> = [a3r] as in *care* vs. <arr> = [a1r] as in *carry*

<erV> = [e3r] as in *here* vs. <err> = [e1r] as in *herring*

<irV> = [i2r] as *mire* vs. <irr> = [i1r] as in *mirror*

<orV> = [o5r] as in *bore* vs. <orr> = [o1r] as in *borrow*

The VrV/Vrr contrast is somewhat different with <u>: <urV> = [u3r] as in *sure* or [yu3r] as in *cure* vs. <urr> = [u1r] as in *current*

The pattern does not apply to <yrV>, as in *lyric*, *pyramid* with <yrV> and a short head vowel, or to <yrr>, as in the rare and technical *pyrrrole* with [e3r].

Notice that the Vrr pattern usually involves the regular short vowel with very little, if any, [r]-coloring of the vowel. Among the long vowels, [i2r] also involves very little or no [r]-coloring, [o5r] involves more, [a3r] and [e3r] even more yet.

**CV#** = Consonant + long vowel at the end of the word, stressed or unstressed, as in *by* and *many*

**CV#X** = Consonant + reduced vowel at end of word, as in *larva*

**CVC#** = Consonant + stressed short vowel + consonant at end of the word, as in *bat* (AES, 93-94)

**CVC#X** = Holdout to CVC#, as in *control*

**V.V** = Long vowel+vowel with syllable boundary between them, as in *lion* (AES, 91-93)

**VCCle** = Stressed short vowel+consonant+consonant+<le>, as in *little* and *candle* (AES, 105-06)

**VCle** = Stressed long vowel+consonant+<le>, as in *title* (AES, 105-06)

**VCleX** = Holdout to VCle, as in *butler* (AES, 105-07)

**VCr** = Stressed long vowel+consonant+<r>, as in *secret* (AES, 106)

**VCrX** = VCr holdout, as in *fabric*

**FLR** = Instance of French Lemon Rule with short head vowel in a VCV string, as in *dragon* (AES, 127-28, where it is called the Stress Frontshift Rule). In addition to its influence on disyllables from French, due to the influence of analogy, it also affects longer words derived from these French

disyllables: *scholar, scholarly, scholarliness*, etc. Additionally, it affects some three syllable words from French and their longer derivations: *consider, consideration; continue, continually*, etc.

**3VR** = Instance of Third Vowel Rule with short head vowel in a VCV string, as in *national*. Sometimes the vowel in question is more than the third vowel from the end, as in *nationalization*. (AES, 131-142, where it is called the Third Syllable Rule)

**ITY** = Instance of the Suffix *-ity* Rule, as in *sanity*, with short head vowel in VCV (AES, 112-15)

**IC** = Instance of the Suffix *-ic* Rule, as in *critic*, with short head vowel in VCV (AES, 115-18)

**ICX** = Holdout to the Suffix *-ic* Rule, as in *aerobic* (AES, 116-18)

**ION** = Instance of the Suffix *-ion* Rule (AES, 118-19), as in *addition*, and *agitation* and *conclusion*, with short <i> or long <a, e, o, u> as head of VCV string (AES, 118-19)

**IONX** = Holdout to Suffix *-ion* Rule, as in *companion*

**IT** = Instance of the Suffix *-it* Rule (AES, 120), as in *credit*

**ITX** - Holdout to Suffix *-it* Rule, as in *unit*

**cb** = Contains a consonant blend, like the <nt> in *agent*

**cd** = Contains a consonant digraph, as in *with* (AES, 71-72).

**cs** = Contains a consonant simplification – two or more letters that spell a single consonant sound due to sound change, as at the end of *bomb*.

**ct** = Contains a consonant trigraph, as in *witch*

**db** = Contains a consonant doublet or doublet equivalent, as in *kiss* and *scene*.

**vd** = Contains a vowel digraph, as in *head*. Spellings of diphthongs are tagged as vowel digraphs.

**vt** = Contains a vowel trigraph, as in *ambitious*

**CMP** = Compound word, as with *baseball*

**DELE** = Instance of silent final <e> deletion, as in *devotion* devoté+ion] (AES, 145-60)

**DEL!** = Nonregular final <e> deletion, as in *argument* argué+ment] and *pastry* pasté+ry] (AES, 158-59) or other unusual deletions.

**PELE** = Instance of penultimate <e> deletion, as in *angry*, angér+y]

**TR** = Contains an instance of twinning or has derived forms with twinning, as in *batter* or *bat* (AES, 161-76)

**ASSIM** = Assimilation of final consonant in a prefix, as in *co+n+cert* (AES, 177-98)

**EXS** = Deletion of <s> after prefix *ex-*, as in *expect*, [ex+\$pect

**Themes.** In this field more than 7560 words are tagged for 169 themes, or topics, with which they can be associated. It is intended to be useful for generating word lists dealing with a common theme, such as “Colors” or “Sports”. There is nothing very authoritative or exhaustive about these taggings. Subjective judgements abound, and occasional violence is done to some formal, scientific categories. All I can say is that on at least one day, one retired English teacher saw each word belonging to the various

themes for which it was tagged. Due to homography, as a given form moves from one theme to another, it often becomes a different word. For instance, the form <molar> in the Science4 theme is a homograph of the form <molar> in the Anatomy1 theme – that is, an entirely different word with the same spelling.

The following is a full list of the themes, many of which are organized into groups, tagged with a group name and a numerical index: **Anatomy1**, **Anatomy2**, etc. The description of each theme concludes with a parenthesis containing two example words and the number of words tagged for that theme. Fields can be combined so that if you want words that are appropriate for first and second graders and that deal with, say, insects, you would filter on “Theme contains Animals2” and “Rank contains AA.” The same strategy can be used to generate lists of, say, nouns about family members for second graders: “Theme contains Family1” and “Rank contains A” and “Part of speech contains \*s\*” (for “substantive”) returns a list of nineteen family nouns.

A more compact and orderly presentation of the following information can be found in the data table titled “Themes”, with separate fields for the eight grade levels, substantives (S), adjectives (J), verbs (V), adverbs (B), and prepositions (E).

**Anatomy1** lists words dealing the skeletal and muscular systems of the body (*ankle, vertebrate*; 112).

**Anatomy2** lists words dealing with the organs, genes, and glands (*skin, gastrointestinal*; 99).

**Anatomy3** lists words dealing with fluids and other substances within the body (*blood, insulin*; 33).

**Animals1** lists birds and things associated with birds (*crow, hatch*, 52);

**Animals2** does the same for insects (*beetle, hive*, 34),

**Animals3** for warm-blooded animals, (but not birds) (*kangaroo, hominid*, 106),

**Animals4** for cold-blooded reptiles and fish, and a few others that actually do not contain blood, like mollusks and sponges (*oyster, invertebrate*, 34).

**Animals5** lists miscellaneous words dealing with animals in general (*hibernate, zoo*, 31).

**Archaic** lists words that were common in earlier English but are now encountered mostly in early texts, such as the King James Bible (*couldst, spake*, 21).

**Art1** lists words dealing with print and literature (*haiku, manuscript*, 132);

**Art2** lists words referring to stage and film art (*actress, playwright*, 63);

**Art3**, words referring to musical instruments and voices (*baritone, guitar*, 50);

**Art4**, words dealing with musical types and qualities (*jazz, allegro*, 68);

**Art5**, words dealing with the visual arts: painting, sculpture, architecture, etc. (*impressionism, architecture*, 94);

**Art6**, words dealing with miscellaneous aspects of the world of music (*Beethoven, octave*, 68).

The Business group deals with the world of business and commerce:

**Business1**, words for grades one and two (*market, shop, 74*);

**Business2**, words for grade three (*capital, credit, 76*);

**Business3**, grade four (*career, employer, 85*);

**Business4**, grades five and six (*international, promotion, 128*);

**Business5**, grades seven and eight, including several words from Hirsch et al's *Dictionary of Cultural Literacy* (*bonus, inventory, 79*).

**Business6**, more business and commerce words from Hirsch (*bureaucrat, 45*).

**Calendar** lists the names of months, weekdays, and holidays, as well as periods in the day (*Friday, Halloween, 54*).

**Cities** lists the names of cities of the world (*Venice, Philadelphia, 46*).

**Clothing1** lists specific articles of clothing (*hat, shirt, 44*).

**Clothing2** lists accessories and parts of clothing (*collar, lace, 50*).

**Colors** lists the names of colors and their qualities (*green, bright, 53*).

**Communication1** lists verbs that deal with the various functions or uses of communication acts (*announce, persuade, 92*).

**Communication2** lists nouns that refer to various products or end results of communication (*argument, understanding, 69*).

**Communication3** deals with miscellaneous methods, aspects, and qualities of communication (*media, conciseness, 93*).

**Containers** lists various kinds and attributes of containers (*box, enclose, 54*).

**Countries** lists proper nouns that name countries (*Japan, Italy, 39*).

**Crime1** lists types of crime and criminal (*embezzlement, murderer, 69*).

**Crime2** lists words about various people and things involved in the law and justice system (*police, court, 101*).

**Crime3** a miscellaneous group of things and qualities involved with crime in general (*contraband, offense, 75*).

**Entertainment1** lists types of entertainment and entertainers (*magician, roulette, 69*).

**Entertainment2** lists actions of people who are being entertained and the effects entertainment has on them (*mirth, excitement, 65*).

**Entertainment3** lists miscellaneous words that refer to entertainment in one way or another (*costume, audience, 101*).

**Family1** lists members of a family (*parent, sister, 67*).

**Family2** lists actions, events, qualities, and things relating to families (*birthday, inheritance, 85*).

**Farming1** lists things raised on farms (*crops, mutton, 64*).

**Farming2** lists equipment and workers found on farms and the things they do (*tractor, pesticide, 53*).

**Farming3** lists words that refer to miscellaneous things related to farming (*meadow, graze, 59*).

**Feeling1** lists nouns that refer to positive feelings (*confidence, vigor, 79*).

**Feeling2** lists nouns that refer to negative feelings (*fright, woe, 88*).

**Feeling3** lists positive adjectives (*beautiful, fearless, 106*).

**Feeling4** lists negative adjectives (*greedy, treacherous, 95*).

**Feeling5** lists positive verbs (*enjoy, trust, 43*).

**Feeling6** lists negative verbs (*distress, vex, 65*).

**Feeling7** lists miscellaneous words related to various aspects of feelings, including a number of adverbs (*happily, wish, 97*).

**Food1** lists words dealing with fruits and nuts (*strawberry, walnut, 44*).

**Food2**, lists grains and bread (*wheat, biscuit, 32*).

**Food3** lists meat, fish, poultry, and dairy products (*beef, egg, 60*).

**Food4** lists sweets (*cookies, pudding, 32*).

**Food5** lists vegetables (*turnip, potato, 32*).

**Food6** lists drinks (*juice, pop, 50*).

**Food7** lists miscellaneous words dealing with food (*eating, buffet, 90*).

The Gender group lists words dealing with gender, sex, and sex difference:

**Gender1** lists words that mark the distinction between male and female for people and other creatures (*son, daughter, 83*).

**Gender2** lists words dealing with sex and reproduction (*conception, penis, 52*).

**Gender3** lists words dealing with miscellaneous aspects of gender and sexuality (*herpes, sexism, 39*).

**Geography1** lists common and proper nouns referring to geographical places, excluding countries and cities (*Europe, planet, 28*).

**Geography2** lists nouns appropriate for grades one through four that refer to natural geographical features (*desert, ocean, 53*).

**Geography3** lists nouns for grades five and up that refer to natural geographical features (*bayou, isthmus, 63*).

**Geography4** lists miscellaneous words dealing with geography (*environment, geology, 73*).

**Government1** lists nouns that refer to people and groups involved in the governing process (*king, congress, 111*).

**Government2** lists mostly abstract nouns (and a few modifiers) that are appropriate for grades one through six and refer to types of government, their aspects and qualities (*democracy, rights, 52*).

**Government3** lists more abstract nouns and modifiers that are appropriate for grades seven and eight (*communism, impeachment, 92*).

**Government4** lists words that deal with the process of governing (*campaign, filibuster,*

90).

**Government5** lists verbs that refer to actions of governments (*appoint, install*, 38).

**Government6** lists miscellaneous words dealing with government and governing (*bandwagon, gerrymander*, 102).

**Groups1** lists words that refer to groups that always, or at least usually, contain people (*junta, panel*, 82).

**Groups2** lists words that refer to all other kinds of groups (*bunch, litter*, 97).

The Health group includes words dealing with health, sickness, and death:

**Health1** lists words dealing with medications and drugs (*antibiotic, cortisone*, 36).

**Health2** lists words dealing with care and treatment (*dentist, hospital*, 87).

**Health3** lists nouns that refer to strictly or mostly mental conditions (*hysteria, phobia*, 50).

**Health4** lists words dealing with physical illness, diseases, and death (*asthma, cardiac*, 113).

**Health5** lists adjectives dealing with health (*mortal, tender*, 67).

**Health6** lists verbs (*relieve, suffer*, 63).

**Health7** lists nouns (*calorie, injury*, 81).

**History1** lists nouns and adjectives dealing with American history (*colonial, Lincoln*, 72).

**History2** lists nouns and adjectives dealing with ancient history (*classical, Troy*, 30).

**History3** lists those dealing with European history (*knight, Napoleon*, 67).

**Home1** lists movable furniture and furnishings found in the home (*couch, blanket*, 68).

**Home2** lists fixtures, rooms, and spaces (*ceiling, den*, 68).

**Home3** lists miscellaneous words associated with house and home (*address, deed*, 105).

**Language1** lists words dealing with: grammar, spelling, word structure, parts of speech, and punctuation (*alphabet, noun*, 102).

**Language2** lists words dealing with semantics and meaning (*dictionary, meaning*, 34).

**Language3** lists words dealing with the spoken language and pronunciation (*homophone, pronounce*, 37);.

**Language4** lists words dealing with rhetoric, or the uses of language and its effects (*argument, slang*, 57).

**Language5** lists miscellaneous words dealing with language, including the names of various languages (*Japanese, printing*, 42).

**Light1** lists verbs about light and its qualities (*gleam, reflect*, 39).

**Light2** lists nouns (*moonlight, sheen*, 34).

**Light3** lists adjectives and adverbs (*brilliant, intense*, 24).

The Location group includes locations, positions, and directions:

**Location1** lists words that are either prepositions or adverbs or both and are appropriate for grades one and two (*beyond, under, 61*).

**Location2** lists prepositions or adverbs appropriate for grades three and four (*opposite, wherever, 35*);

**Location3** lists prepositions or adverbs appropriate for grades five through eight (*offshore, underground, 34*).

**Location4** lists other location words appropriate for grades one through three (*corner, middle; 53*);

**Location5**, appropriate for grades four through six (*latitude, suburb, 44*);

**Location6**, appropriate for grades seven and eight (*perigee, longitude, 15*).

**Materials1** lists metals and metallic materials (*alloy, wire, 24*).

**Materials2** lists minerals and mineral-like materials (*coal, pearl, 52*);

**Materials3** lists materials from vegetable matter or from animals, including from petroleum (*charcoal, leather, 62*).

**Materials4** lists miscellaneous words that deal with materials and are hard to fit into any of the preceding three (*stuff, plastic, 30*).

**Math1** lists number names (*digital, sixteen, 74*)

**Math2** lists mathematical concepts and calculations (*equal, multiply, 59*).

**Math3** lists miscellaneous math words appropriate for grades one through four (*pair, problem, 19*).

**Math4** lists miscellaneous words for grades five and six (*plus, subset, 31*).

**Math5** lists miscellaneous words for grades seven and eight, including several from Hirsch *et al* (*axiom, linear, 65*).

**Measure1** lists adverbs dealing with amounts, degree, and sizes (*often, loudly, 64*).

**Measure2** lists adjectives (*enormous, abundant, 82*).

**Measure3** lists nouns (*amount, handful, 75*).

**Measure4** lists words dealing with calculated measurements (*breadth, frequency, 79*).

**Measure5** lists units of measurement, including monetary units (*dollar, inning, 96*).

**Military1** lists words about military paraphernalia and equipment (*helmet, rocket, 53*).

**Military2** lists military personnel (*captain, regiment, 80*).

**Military3**, lists military actions and operations (*raid, maneuver, 80*).

**Military4** lists miscellaneous military words (*honorable, strategic, 72*).

**Mind1** lists nouns (mostly rather advanced) that refer to types and schools of intellection or thought (*philosophy, science, 51*).

**Mind2** lists verbs referring to various mental acts (*believe, calculate, 76*).

**Mind3** lists nouns referring to the results of mental acts (*discovery, certainty, 142*).

**Mind4** lists words dealing with miscellaneous aspects of the mind and mental acts, including psychological constructs (*meanings, irrational, 57*).

**Myth** lists names and qualities of myths and mythological figures, ancient and modern

(*phoenix, werewolf*, 56)

**Names** lists common names, both first names and last (*Dorothy, Franklin*, 88).

**Occupation1** lists nouns with the agent suffixes *-ar]2*, *-er]01*, *-or]2*, or *-ess]1* (*advisor, farmer*, 79).

**Occupation2** lists other occupational nouns (*housewife, politician*, 99).

**People1** lists nouns and pronouns appropriate for grades one and two that refer to individual people – their roles, jobs, demeanors (*group, officer*, 91).

**People2** lists such words for grade three (*chum, nurse*, 101).

**People3** lists such words for grade four (*bride, follower*, 181).

**People4** lists such words for grades five and six (*fugitive, magician*, 258).

**People5** lists words for grades seven and above, including words from Hirsch *et al.* (*baritone, extrovert*, 150).

**Plants1** lists nouns and adjectives appropriate for grades one through four that refer to or describe plants and their parts (*cotton, leaves*, 64).

**Plants2** lists nouns and adjectives for grades five through eight (*deciduous, lavender*, 57).

The Religion group lists words dealing with various aspects of religion and religions, including Christianity, Judaism, Islam, Buddhism, Hinduism, and related areas.

**Religion1** lists words appropriate for grades one through four that deal with various aspects of religions and religiousness (*blessing, faith*, 103).

**Religion2** lists such words for grades five and six (*sermon, eternity*, 88).

**Religion3** lists words from for grades seven and above, including words from Hirsch *et al.* (*godliness, heretic*, 124).

**School1** lists verbs dealing with school and schooling (*read, subtract*, 64).

**School2** lists nouns, and a few adjectives, appropriate for grades one through four (*class, excellent*, 83).

**School3** lists nouns, and a few adjectives, appropriate for grades five through eight, including words from Hirsch *et al.* (*fraction, calculator*, 70).

**Science1** lists nouns referring to kinds of science, and a few quasi-sciences, (*chemistry, alchemy*, 36).

**Science2** lists nouns and adjectives dealing with biology and appropriate for grades one through four (*feather, gene*, 59).

**Science3** lists biological nouns and adjectives for grades five through eight (*bacteria, evolution*, 150).

**Science4** lists nouns and adjectives for chemistry (*element, oxygen*, 91).

**Science5** lists nouns and adjectives for physics and astronomy (*comet, particle*, 152).

**Science6** lists verbs, and a few nouns, referring to scientific actions and processes (*analyze, experiment, 63*).

**Science7** lists miscellaneous nouns and adjectives dealing with science and technology (*laboratory, formula, 81*).

**Senses1** lists words dealing with speech and hearing (*ear, loud, 63*).

**Senses2** lists words dealing with sight (*look, visible, 30*).

**Senses3** lists words dealing with smell and taste (*nostril, sweet, 19*).

**Senses4** lists miscellaneous words dealing with other senses and senses in general (*extrasensory, pain, 46*).

**Sports1** lists sports equipment (*diamond, ball, 64*).

**Sports2** lists other sports words appropriate for first and second grades (*game, race, 70*).

**Sports3** lists sports words for third grade (*league, underdog, 70*).

**Sports4** lists words for fourth grade (*eagle, surf, 73*).

**Sports5** lists words fifth and sixth grades (*handicap, skiing, 75*).

**States** lists the names of states (*Michigan, Oregon, 51*).

**Time1** lists time words for grades one and two (*hour, soon, 59*).

**Time2** lists words for grades three and four (*clock, sunset, 61*).

**Time3** lists words for older students (*eternity, prompt, 64*).

**Tools1** lists words about tools (defined rather broadly) for grades one through three (*jack, phone, 42*).

**Tools2** lists tool words for grades four through eight (*computer, hydraulic, 77*).

**Transportation1** lists words dealing with vehicles and other means of transportation (*bicycle, shuttle, 52*).

**Transportation2** lists words dealing with routes, roads, times, and places (*interstate, arrival, 51*).

**Transportation3** lists other transportation words appropriate for grades one through four (*freight, passenger, 66*).

**Transportation4** lists words for grades five through eight (*gasohol, supersonic, 43*).

**Trees** lists the names and other features of trees (*birch, forest, 54*).

**Value1** lists nouns that refer to qualities to which we ascribe subjective values, good or bad. (*curse, friendship, 169*).

**Value2** lists value-laden adjectives (*excellent, ugly, 173*).

**Value3** lists value-laden verbs (*forgive, pollute, 109*).

**Weather1** lists nouns that refer to weather and climate (*cloud, meltdown, 82*).

**Weather2** lists weather and climate adjectives (*dusty, tropical, 27*).

**Homophones.** Homophones are words that sound the same but mean different things and are spelled differently, as with *pear* and *pare*. They can pose special problems for spellers and can benefit from some special attention.

**Homographs.** Homographs are words that are spelled the same but that mean different things and usually are pronounced differently. They pose no particular problems for spellers, but they can for readers. Many of them contrast in pronunciation simply by shifts of stress and contrast in meaning simply by shifts in part of speech – for instance, *convict*, a noun with stress on the first syllable vs. *convict*, a verb with stress on the second syllable.

**Other Problem Spellings.** This field is a companion to Homophones, covering a variety of problems that could benefit from some special attention. It lists near homophones and non-homophonic look-alike words – such as *accept* vs. *except*, *latter* vs. *later*, and *angle* vs. *angel*. Common misspellings are tagged with an asterisk. Words tagged with an exclamation point appear on at least one list of spelling demons for youngsters.

**Spelling Difficulty.** This field lists the level of difficulty for nearly half of the words in CommonWords, based on the percentages of fourth graders who spelled the given word correctly in *The New Iowa Spelling Scale* (Iowa City: State University of Iowa, n.d.). A suggested categorization would be:

- 1-4 = Very hard (360 words)
- 5-13 = Hard (795)
- 14-47 = Medium (1504)
- 48-71 = Easy (755)
- 72-99 = Very Easy (382)

**Rank.** This field is also meant to help in deciding when to introduce certain words to students. It is generally based on the Thorndike-Lorge *Teacher's Word Book of 30,000 Words* (New York: Teachers College Press, 1944, 1972) (T-L), which is primarily aimed at readers rather than spellers. The T-L score used here is that given in the "G" column in their list, which gives the number of occurrences per one million running words. T-L suggest appropriate grade levels. "A" would include "AA"; "B" would include "A" and "AA", etc.:

- AA = Appropriate for grades 1-2.
- A = Appropriate for grade 3
- B = Appropriate for grade 4 (A T-L score of 49-20)
- C = Appropriate for grades 5-6 (A T-L score of 19-10)
- D = Appropriate for grades 7-8 (A T-L score of 9-1)

For more on T-L's rankings see T-L, pp. x-xii. Words with a T-L score between 1 and 6 or that have no T-L score are assigned to a grade level based on my informed best

guess, supported with the rankings in the *The American Heritage Word Frequency Book*. The original T-L scores are based strictly on frequency; my assignments try to balance frequency with difficulty. Obviously these assignments are quite approximate.

T-L normally do not list inflected forms separately. The score they list for the base form sums up all of the inflected forms. Since CommonWords does list many inflected forms separately, I've chosen usually to give the inflected forms the same ranking as that of the base form listed in T-L. Exceptions to this procedure are cases where there is a complication in the spelling of the inflected form (that is, a deletion, a twinning, or a change of <y> to <i> or of <i> to <y>), in which cases I've adjusted the ranking of the inflected form up one level so that the inflected form of "AA" words becomes "A", and those of "A" words become "B". I did not make this adjustment on words ranked "B" or higher.

The 2000+ words marked "H" in the Rank field are a special group. Those marked simply "H" do not appear in T-L's main word list, but are drawn from Hirsch *et al's Dictionary of Cultural Literacy*. Words from Hirsch that occur in T-L are tagged with their normal T-L score; all of those with a T-L score between 9 and 1 are tagged "DH". The words from Hirsch play a particularly important part in what Hirsch and his team call cultural literacy, the "common knowledge or collective memory [that] allows people to communicate, to work together, and to live together". Obviously, the same could be said for all of the words in CommonWords (and many, many others), but the words tagged "H" have a special importance. They are only a sample, for the *Dictionary of Cultural Literacy* includes thousands of other words and phrases, including many proper names of people, places, events, and things that are for the most part excluded from CommonWords. I have included these "H" words because I believe it is important for students to be exposed to such words as soon as possible, even though they are often quite technical and advanced.

**Range and Subrange.** The Range field indicates into which of five ranges each of 5680 tagged words falls. Ranges are intended to provide help in finding words appropriate to the students' level of mastery. For instance, the 1,000 plus words in Range 1 are all completely regular and completely analyzable if the students have had work with the Range 1 sound-to-spelling correspondences, which are listed below. The ranges are organized so that each of the first four ranges contains only one spelling for each sound and only one sound for each spelling. This regularity is not true of the correspondences in Range 5, due to the existence of several sounds that have more than five different spellings.

Subranges 1a, 1b, 2a, and 2b are subsets of ranges 1 and 2. Subrange 1a consists of Range 1 words that contain only the consonant and short vowel correspondences from Range 1. It contains words with the regular patterns for short vowels – namely, VCC and VC#. Subrange 1b consists of words that contain only the consonant and long vowel correspondences from Range 1, and the regular patterns for long vowels – VCe#, VCV, and several digraphs. Subrange 2a consists of Range 2 words that contain only the Range 1 and 2 consonant and short vowel correspondences. Subrange 2b consists of words that contain only Range 1 and 2 consonant and long vowel correspondences.

The Range 1 correspondences are these 35:

**The Short Vowels:**

[a1] = <a> as in *pat*  
[e1] = <e> as in *pet*  
[i1] = <i> as in *pit*  
[o1] = <o> as in *pot*  
[u1] = [u] as in *but*

**The Long Vowels and Diphthongs:**

[a2] = <a...e> as in *mate*  
[e2] = <ee> as in *meet*  
[i2] = <ie> and <i...e> as in *pie* and *pile*  
[o2] = <oe> and <o...e> as in *woe* and *quote*  
[u2] = <oo> as in *boot*  
[yu2] = <ue> and <u...e> as in *hue* and *huge*  
[oi] = <oi> as in *foil*  
[ou] = <ou> as in *foul*

**The Consonants:**

[b] = <b> as in *bob*  
[d] = <d> as in *dad*  
[f] = <f> as in *fluff*  
[g] = <g> as in *gag*  
[h] = <h> as in *hot*  
[j] = <j> as in *jot*  
[k] = <c> as in *cat*  
[l1] = <l> as in *lot*  
[m] = <m> as in *mom*  
[n1] = <n> as in *nun*  
[ng] = <ng> as in *bring*  
[p] = <p> as in *pop*  
[r] = <r> as in *roar*  
[s] = <s> as in *sit*  
[t] = <t> as in *tot*  
[v] = <v> as in *vine*  
[w] = <w> as in *wine*  
[y] = <y> as in *yet*  
[z] = <z> as in *zip*  
[ch] = <ch> as in *chin*  
[sh] = <sh> as in *shin*  
[th1] = <th> as in *thin*

This may seem like a lot of correspondences, but notice that in nearly every case the spelling uses the same letter as we normally use to symbolize the sound. The symbol "...e>" indicates that the long vowel letter is followed by a single consonant letter and a silent final <e>, which is marking the long vowel sound. Most of these correspondences

are very high frequency. Vowels that precede [r] often vary considerably in their pronunciation from that when they precede some other consonant. Consider, for instance, the different pronunciations of <a> in *mare* and *mate*.

**Range 2.** The 800+ Range 2 words are completely regular and analyzable if the students have had work with the Range 1 correspondences and the following 33:

### **The Short and Reduced Vowels:**

[e1] = <ea> as in *bread*

[i1] = <e> as in *basket*

[o1] = <a> as in *ball*

[u1] = <o> as in *from*

[u3] = <oo> as in *wood*

[u4] (schwa) = <a> as in *allow*

### **The Long Vowels and Diphthongs:**

[a2] = <ai> as in *rain*

[e2] = <e...e> as in *theme*

[i2] = <y...e> as in *type*

[o2] = <oa> as in *boat*

[u2] = <ue> and <u...e> as in *due* and *dune*

[yu2] = <ew> as in *few*

[oi] = <oy> as in *coy*

[ou] = <ow> as in *coil*

[a3r] = <air> as in *hair*

[o3r] = <or> as in *cord*

### **The Consonants:**

[b] = <bb> as in *ribbon*

[d] = <dd> as in *ridden*

[f] = <ff> as in *stuff*

[g] = <gg> as in *rugged*

[j] = <g> as in *large*

[k] = <k> as in *lake*

[l1] = <ll> as in *tall*

[m] = <mm> as in *summer*

[n1] = <nn> as in *runner*

[ng] = <n> as in *brink*

[p] = <pp> as in *happy*

[r] = <rr> as in *marry*

[s] = <c> as in *cent*

[t] = <tt> as in *attic*

[w] = <u> as in *quit*

[y] = <i> as in *onion*  
[z] = <s> as in *dogs*  
[ch] = <tch> as in *catch*  
[sh] = <s> as in *sure*  
[th2] = <th> as in *then*

It would be good, though not necessary, for the students to have worked with the reasons for double consonant letters: twinning, the assimilation of consonants at the end of prefixes, simple addition, and the VCC tactical pattern.

**Range 3.** The 1,000+ Range 3 words are completely regular and analyzable if the students have had work with Ranges 1 and 2 and with the following correspondences and tactical patterns:

#### **The Vowels:**

[a1] = <au> as in *laugh*  
[i1] = <y> as in *system*  
[o4] = <aw> as in *law*  
[u3] = <u> as in *put*  
[a2] = <ay> as in *day*  
[e2] = <ea> as in *speak*  
[o2] = <ow> as in *low*  
[u2] = <o...(e)> as in *move*  
[yu2] = <eu> as in *feud*  
[u4] = <e> as in *children*  
[u4r] = <er> as in *batter*

#### **The Consonants.**

[f] = <gh> as in *laugh*  
[h] = <wh> as in *whole*  
[j] = <d> as in *graduate*  
[k] = <ck> as in *pick*  
[r] = <wr> as in *write*  
[s] = <ss> as in *miss*  
[z] = <zz> as in *buzz*

In addition to these sixteen correspondences Range 3 words assume that the students have had work with two tactical patterns for long vowels: (i) the stressed head vowels of VCV strings are normally long – for instance, the <a> in *bacon* spells [a2] , and (ii) vowels at the end of syllables are also regularly long – for instance, the <i> in *lion* spells [i2]. The first of these two, which is essentially an extension of the Range 1 and 2 correspondences with “...e>”, is discussed in *AES* as the VCV pattern, the second as the V.V pattern.

**Range 4.** The 1,000+ Range 4 words are completely regular and analyzable if the students have had work with Ranges 1, 2 and 3 and with the following correspondences and tactical patterns:

**The Vowels:**

[i1] = <a> as in *chocolate*

[o5r] = <ar> as in *hard*

[o4] = <au> as in *sauce*

[u] = <oo> as in *blood*

[a2] = <ea> as in *break*

Weak [e2] = <y> as in *funny*

Strong [e2] = <ie, ei> as in *piece, receive*

[u2] = <ew> as in *drew*

[u4] = <io> as in *region*

[u4l] = <le> as in *jungle*

[u4r] = <or> as in *doctor*

[yu4] = <u> as in *deputy*

[yu3r] = <ur...(e)> as in *cure*

**The Consonants:**

[f] = <ph> as in *telephone*

[j] = <dg> as in *judge*

[ks] = <x> as in *fix*

[k] = <q> as in *quit*

[n1] = <kn> as in *know*

[r] = <rh> as in *rhythm*

[s] = <sc> as in *scene*

[sh] = <t> as in *nation*

In addition to these eighteen correspondences Range 4 words assume that the students have worked with silent final <e>'s that serve various diacritical functions other than marking long vowels and with silent final <e>'s that serve no diacritical function at all. It also assumes familiarity with the <i>-before-<e> pattern. Holdouts to this pattern with <ei> are included in Range 5.

**Range 5.**

**The Vowels.**

[a3r] = <are> as in *rare*

[a1r] = <ar> as in *tariff*

[a1r] = <arr> as in *carriage*

[e2] = <ei> not after <c> as in *neither*

[e2] = <i> as in *machine*  
[u4] = <i> as in *horrible*  
[u4] = <o> as in *million*  
[u4] = <u> as in *awful*  
[u4] = <ou> as in *courteous*  
[u4r] = <ar> as in *coward*  
[u4r] = <ur> as in *injury*

### **The Consonants.**

[gz] = <x> as in *exact*  
[k] = <cc> as in *account*  
[k] = <ch> as in *school*  
Syllabic [l] = <l> as in *battle*  
[u1r] = <ear> as in *earth*  
[u1r] = <er> as in *term*  
[u1r] = <ir> as in *firm*  
[u1r] = <our> as in *courage*  
[u1r] = <ure> as in *sure*  
[t] = <ght> as in *night*  
[hw] = <wh> as in *why*  
[ch] = <t> as in *feature*  
[sh] = <c> as in *social*  
[sh] = <ss> as in *mission*  
[zh] = <s> as in *casual*

Range 5 words also assume some work with the VCle# long vowel pattern, with the apostrophe, and with non-diacritical, non-final silent <e>'s.

**Characters.** The number of characters (letters, punctuation marks, and blank spaces) in each word.

**Syllables.** The number of syllables in each word. For some words the *AHD* shows variant pronunciations with different numbers of syllables – for instance, one pronunciation of *average* has three syllables, another has only two, so the Syllables field shows both: “3 2”. A few final syllables are quite weak, consisting of only a syllabic consonant, as in *button* and *little*.

**Syllable Structure.** This field is for people who work with the notion of closed vs. open syllables and with word stress. Closed syllables, ending with a consonant sound, are tagged C; open syllables, ending with a vowel sound, are tagged O. Additionally, unstressed, syllables are tagged “u”; stressed, syllables are tagged “s”. The vowel in each syllable is tagged (i) “t” if it is tense – that is, in general, orthographically long, (ii) “l” if it is lax, or orthographically short, or “r” if it is reduced to schwa or is spelled with a syllabic consonant. Thus, the tagging for the word *adorn* is “OurCst”, which means

that the first syllable is open, unstressed, with a reduced vowel, while the second syllable is closed, stressed, and tense. The tagging of *sequence*, “OstCul”, means that the first syllable is open, stressed, and tense (or long), while the second syllable is closed, unstressed, and lax (or short).

The following are some useful filters for stress patterns:

For the “iambic” stress pattern (that is, unstressed-stressed): “%u%s%”

For the “trochaic” stress pattern (that is, stressed-unstressed): “%s%u%”

For the “spondaic” stress pattern (that is, stressed-stressed): “%s%s%”

Primary and secondary stress are not distinguished here, both being represented with a simple “s”. Several words have more than one stress pattern, depending usually on the part of speech they are filling – for instance, the iambic verb *convict* vs. the spondaic noun *convict*. Also, to avoid getting false hits, it’s a good idea when searching this field to indicate in the Syllables field the length of the words in which you are interested.

Unstressed short <i> I treat as lax though it could legitimately be treated as reduced. For one thing, it often occurs in open syllables like schwa and unlike stressed lax vowels.

**Complexity.** *Complexity* here means the number of different orthographic features listed in the Analysis field. Words tagged 4-7 in this field have four or more features and are thus orthographically the most complex; those tagged 2-3 have two or three; those tagged 0-1 have one or none.

**Prefixes and Suffixes.** The Prefixes and Suffixes fields provide a very large sample of the prefixes and suffixes that can be affixed to each word in CommonWords following the normal procedures of combination – such as simple addition, final <e> deletion, twinning final consonants, <y> to <i> replacements, and occasional vowel deletions to avoid unwanted double vowels – as with the <a> deletion when *-an1* is affixed to *utopia*. These fields offer just a sample, with no attempt at exhaustivity. Actually, exhaustivity would not be possible, since prefixes and suffixes constitute a huge source of potential new words, only some of which have been put to use and recorded in dictionaries – and many more of which have been put to use as nonce words but never recorded in dictionaries. Combined, the lists in Prefixes and Suffixes bring the total number of words in CommonWords to well over 100,000 – some of which, of course, are not all that common.

I have tried to exclude affixations that produce words identified as archaic, obsolete, or dialectal. Though the affixations given here usually reflect the historical growth of words, in some cases rather than reflecting that history, they are meant simply to suggest relationships of meaning and logic among English words. By and large, the sometimes different explications given in the Lexis database are more consistently historical.

Several of the affixes are numbered to discriminate homographs. To find to which prefix

or suffix each number refers, see the Prefixes and Suffixes tables in the Lexis database. The prefixes and suffixes listed in the drop down menus for these two fields in the user interface are high frequency affixes taken from the Prefixes and Suffixes tables in the Lexis database elsewhere on this site.

**Prefixes.** Prefixes are listed in two places: (i) the Explication field lists any prefixes contained within the listed word itself, tagged with a leading left square bracket; (ii) the Prefixes field lists prefixes that can be added to the listed word.

Actually, the question of what a prefix is remains surprisingly undecided. Dictionaries do not agree on the distinction between prefixes and bases, especially those usually bound bases often called "combining forms." For instance, though the *AHD* uses *electr+* as an example of a combining form (at "combining form"), in the main word list it is labeled "prefix," as are all other combining forms. On the other hand, the *RHUD* and *W3* both distinguish carefully between affixes and combining forms. At a different extreme the editors of *Prefixes: and Other Word-Initial Elements of English* collapse the distinction completely, speaking only of "word-initial elements" in their list of nearly 3,000 forms.

Elements signifying numerical values can illustrate the indecision: In *W3 bi-* "two" is labeled a prefix, but *tri-* "three" is a combining form. *RHUD* labels both as combining forms; *AHD* labels both as prefixes. I treat all numerical elements as combining forms – that is, bases, usually bound – and restrict prefixes essentially to prepositions (*in2-* "in", *ad-* "to, towards"), negatives (*in1-*, *non-*, *un1-*), adverbs (*se-*, *per1-*), and a few derivationals, (*en1-* and *be-*). (For more on this see "Introduction to the Prefixes Table" in the "Introduction to the Lexis Database" article elsewhere on this site. Also the Examples and Comment fields in the drop-down menu are much abbreviated. For more details see the Prefixes table in the Lexis database.)

**Ellipses, Parentheses, Braces, and Diagonals.** In the Prefixes field ellipses represent the word, separating the prefix from any optional or required suffixes. A suffix is optional if the stem formed by the prefix-plus-word is free; a suffix is required if the stem formed by the prefix-plus-word is not free and requires the addition of a suffix to make a recognized word. One or more suffixes enclosed in parentheses are optional and all create recognized words. For example, at *addict*, "non...(ing1 ive ed1)" represents the following: *nonaddict*, *nonaddicting*, *nonaddictive*, *nonaddicted*. However, a single required suffix is not enclosed in parentheses – for instance, at *audit* "un1...ed1" represents only *unaudited*, since we have no recognized word \*unaudit. Two or more suffixes, any one of which are required to make a recognized word, are enclosed in curly braces. Thus, at *admire* "un1...{able ing1}" represents *unadmirable* and *unadmiring* since we do not have the word \*unadmire.

Each of the prefixes enclosed in square brackets can take the suffixes following the ellipses. Thus, at *condense* "[inter ultra]...er01" represents *intercondenser* and *ultracondenser*. Because of space limitations, in a very few cases – at *verse*, for instance – prefixes enclosed in diagonals – as "/con in2 ob re tra un1/" – do not have their acceptable suffixes listed. A prefix enclosed in parentheses and immediately

preceding another prefix can be added to the word formed with the second prefix – thus, at vow “(dis) a3...(al2)” represents *avow*, *avowal*, *disavow*, *disavowal*.

**Suffixes.** Suffixes are listed in three places: (i) the Explication field lists any suffixes contained within the listed word itself, tagged with a following right square bracket; (ii) the Suffixes field lists suffixes that can be added to the listed word, and (iii) in the Prefixes field some suffixes are listed that occur in conjunction with certain prefixes.

In the Suffixes field, suffixes in parentheses can be added only after the immediately preceding suffix has been added. For instance, the word *flesh* can take the comparative suffix *-er]02* “more” only after it has taken the adjective suffix *-y]1*: *fleshier* but not \**flesher* \*”more flesh” – thus, at *flesh*, *y1 (er02)*. I’ve tried regularly to mark with a double right parenthesis the ends of strings containing embedded suffixes. For instance, with the regular <y> to <i> replacements, the word *irony* can take the following suffixes: *ic1 (al1) (ly1) (ness) ist1* – in the words *ironic*, *ironical*, *ironically*, *ironicness*, *ironist*. (N.B. I’ve tried to be consistent and clear in showing the embeddings, but I suspect I have not always succeeded.) .

The main concern here is with derivational suffixes, not inflections. Although regular nouns, verbs, adverbs, and adjectives obviously can add the normal inflectional suffixes, inflectional suffixes are not normally listed in the Suffixes column. I list only those inflections that must be added before adding subsequent derivational suffixes – as with, at *haunt*:: *ed1 (ness)*.

In some cases different suffixes can be affixed to two different senses of a homographic stem. For instance, at the word *camp* the suffixes *-aign* and *-er01* suffixes can only be added to *camp1* with the sense “Field, temporary dwelling” while *-y1* can only be added to *camp2* with the sense “Humorous banality”. I’ve tried neither to include nor exclude all suffixes for all the homographic senses.

Several words can take many different suffixes, so in order to make searching the strings of suffixes easier, I’ve tried consistently to alphabetize the lists.

The Examples and Comment fields in the drop-down menu for this field are much abbreviated. For details see the Suffixes table in the Lexis database.

**Parts of Speech** . This field lists the parts of speech that a word can fill. It uses the following codes:

- ry = Regular adjectives – that is, those that can take the comparative and superlative inflectional suffixes *-er*, *-est* – as in *dark*, *darker*, *darkest* (though in many, or all, cases the comparative and superlative can also be shown periphrastically with *more* and *most*).
- nj = Nonregular adjectives, which includes (i) those that show comparative and superlative only periphrastically, as in *admirable*, *more admirable*, *most*

*admirable*; (ii) those that have comparative and superlative forms with bases different from the positive form, as in *good, better, best*; and (iii) those that only rarely or never have comparative or superlative forms – for instance, ordinals like *eighteenth* ; possessive adjectives like *her, his, your, my, our* ; and certain absolutes like *every, subsequent, prior* .

rb and nb = Regular and nonregular adverbs, similar to the above distinction between regular and nonregular adjectives

rs = Regular substantives – that is, nouns or noun equivalents – that can form plurals with -s or -es – like *cat/cats* or *kiss/kisses*. I use <s>, for *substantive*, to represent nouns and noun equivalents – like the past participle *forbidden*, as in “The forbidden is always tempting.”

ns = Nonregular substantives – that is, those that can form plurals in other ways, including sets like *goose/geese* and *novas/novae*. Several nouns are both regular and nonregular – for instance, *supernova*, which has the plural *supernovas* and the more technical *supernovae*. Also nouns that have the same form for singular and plural are tagged “ns”: *fish, deer*.

rv = Regular verbs—that is, verbs that form the past tense with -ed, like *dress/dressed*

nv = Nonregular verbs – that is, so-called “strong” verbs like *swim/swam* and verbs that have the same form for present and past, like *put*

c = Conjunctions.

p = Pronouns.

e = Prepositions.

a = Articles

tl = Past participles, both regular and nonregular

tn = Present participles

in = Interjections

With several words there is not a perfect match between the analyses in the Sound to Spelling and Spelling to Sound fields and the parts of speech in the Parts of Speech field. For instance, in the Sound to Spelling and Spelling to Sound fields the word *alternate* is analyzed phonetically with a long <a> in the final syllable, which is its pronunciation as a verb. But when *alternate* is used as a noun or adjective, that vowel is destressed to a short <i>. Nevertheless, in the Parts of Speech field *alternate* is tagged as verb, noun, and adjective. One way of thinking about it is that in the phonetics fields we have to settle on one pronunciation, but in the Parts of Speech field we can take an inclusive view, including heterophonic uses of the written word.

**Sources.** The Sources field gives the lineage of each word. A lineage is the language or languages from or through which a word came into English. The immediate source is the last item in the lineage, so complex lineages, which are presented chronologically left-to-right, are most easily read in reverse. Thus the entry “Greek > Latin > French” means “English got the word from French, which got it from Latin, which got it from Greek”. Lineages can get quite lengthy – for instance, that for *sugar* is “Sanskrit > Prakrit > Persian > Arabic > Italian > Latin > French”. (Sanskrit and Prakrit were ancient languages of India.)

CommonWords lineages often simplify the more detailed treatment given in dictionaries. For instance, the etymology given in the *American Heritage Dictionary* for the word *meddle* would lead to the lineage “Latin > Vulgar Latin > Old French > Anglo-Norman”, which is simplified in CommonWords to Latin > French. (Vulgar Latin was the nonliterary, common speech of the Romans. Old French was French as it was spoken from the 9<sup>th</sup> to the 16<sup>th</sup> century. Anglo-Norman was the dialect of Old French spoken by the Normans of Normandy who conquered England in 1066.) Also, CommonWords lineages do not distinguish between chronological periods of a language – for instance, scholars distinguish five ages of Latin: Old Latin (9<sup>th</sup> century B.C. to 3<sup>rd</sup> century B.C.), Latin (3<sup>rd</sup> century B.C. to 2<sup>nd</sup> century A.D.), Late Latin (3<sup>rd</sup> century to 7<sup>th</sup> century), Medieval Latin (8<sup>th</sup> century to 16<sup>th</sup> century), and New Latin (from 16<sup>th</sup> century to the present). In the CommonWords lineages all five ages are collapsed into one, tagged simply *Latin*. The one exception is Old English, which is distinguished from later English. *Norse* refers to any one of the three Scandinavian languages – Swedish, Norwegian, Danish, old and modern. The lineages do not distinguish between High and Low German. The lineages do not always show the etymology of affixes. For instance, *atonement* is tagged “Old English” for the *at* and *one* that form the word, *atone*. But the suffix *-ment*, which is from Latin via French, is not included in the lineage.

“OOO” means “of obscure origin”. A question mark in a lineage usually means “probably” – sometimes “maybe”. Words tagged “Imitative” were usually formed in English, though sometimes it is not clear exactly what is being imitated. Words that come from proper names are tagged “Eponym”. Those few from trademarks are tagged “Trademark”. Those tagged with an exclamation point have etymologies that are surprising or otherwise interesting. Short clues to the stories of these words can be found in the etymology of regular dictionaries; more detail can be found in special etymological dictionaries such as *The Barnhart Dictionary of Etymology* and in the Word History following some words in the *American Heritage Dictionary*; even more detail can be found in books like Charles Funk’s *Thereby Hangs a Tale*.

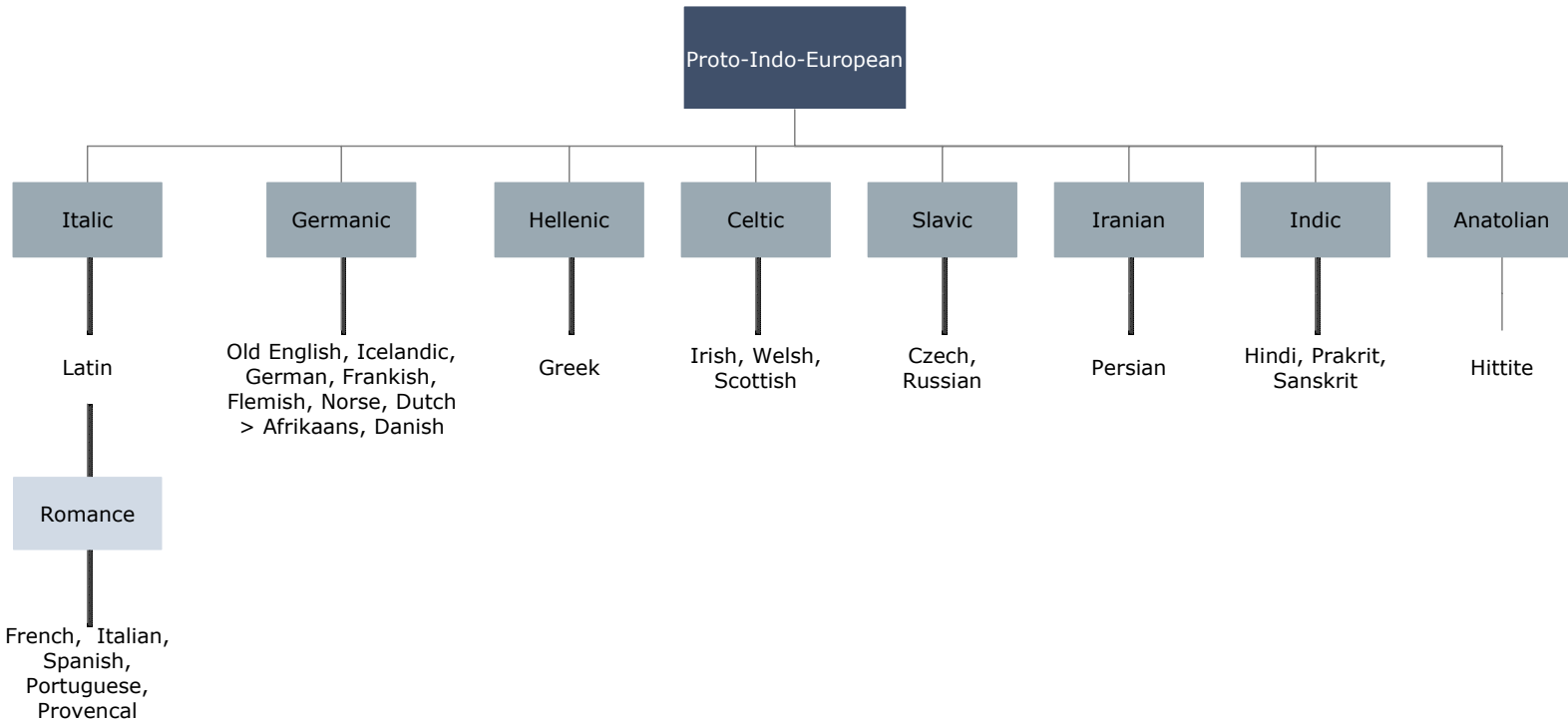
English is one of several languages in the Indo-European super-family, which includes languages in the Slavic, Germanic, Celtic, Italic, Hellenic, Anatolian, and Indic sub-families, and some others not represented in CommonWords. (See the chart below, in which languages appear unboxed, the Romance sub-family in very light gray boxes, families in darker grey, the Indo-European super-family in darkest gray.) Proto-Indo-European, the mother tongue of the Indo-European super-family, is thought to have been spoken around 5000 B.C. in the area north of the Black and Caspian Seas. Over the millennia it is thought to have spread east to India and central Asia, west to modern Greece, Italy, Spain, and north to Germany, Britain, and Scandinavia.

The non-Indo-European Semitic languages descend from the separate super-family, Afro-Asiatic. Semitic languages represented in CommonWords are Canaanite, Akkadian, Arabic, and Hebrew, out of which Yiddish developed.

The tag “Amerindian” includes a number of different non-Indo-European languages from North, Central, and South America. The tag “Tamil” refers to Tamil, a member of the Dravidian language family, spoken in southern India. Sami refers to any of the Finnic

languages spoken by the Lapps.

The Indo-European languages, their families, and a subfamily that are represented in CommonWords:



For suggestions concerning etymological dictionaries see “On Dictionaries and Other Helps for Teaching Vocabulary and Spelling” in the Short Articles venue of this website.

## Other Data Tables:

**Correspondences: Sound to Spelling.** The Correspondences: Sound to Spelling table contains six fields: (i) Sound to Spelling, which lists the sound-to-spelling correspondences; (ii) Examples, which gives an example word for each correspondence; (iii) Instances, which gives the number of words in CommonWords that contain at least one instance of the correspondence; (iv) AES, which cross-references to sections of my *American English Spelling* dealing with the correspondences; (v) Percentages for this Sound, which gives the percentage that this correspondence constitutes for this sound with this spelling, and (vi) Sort, which is a number used to set the sort order for the table. The Sort field can also be used to select subsets of sounds in the correspondences, which are listed in the following order, with the following beginning and ending Sort numbers:

Short (Lax) Vowels: 1-23, 39-44  
Long (Tense) Vowels: 49-93  
Tense but Not long Vowels: 26-38  
Diphthongs: 94-97  
Schwa: 98-117  
[r]-Colored Vowels: 117.1-184  
Vowels with initial [y]: 185-194  
Consonants: 195-302  
Silent Letters: 303-313  
Punctuation: 314-317

The table can be sorted on the Sort and Instances fields. For more information, see the Sounds Count and Spellings Count tables.

**Correspondences: Spelling to Sound.** The Spelling to Sound table contains five fields: (i) Spelling to Sound, which lists the spelling to sound correspondences; (ii) Examples, which gives an example word for each correspondence; (iii) Instances, which gives the number of words in CommonWords that contain at least one instance of the correspondence; (iv) AES, which cross-references to sections of my *American English Spelling* dealing with the correspondences; and (v) Percentages for this Spelling, which gives the percentage that this correspondence constitutes for this spelling. The table can be sorted on the Spelling to Sound and Instances fields. For more information, see the Sounds Count and Spellings Count tables.

**Sounds Count.** This data table lists the frequency of occurrence of each of the 68 sounds found in the 8591 words in CommonWords. It contains nine fields: (i) Sound; (ii) Words with One Instance; (iii) Words with Two Instances; (iv) Words with Three Instances; (v) Words with Four Instances; (vi) Total Instances, which totals the preceding four columns multiplied respectively by 1, 2, 3, or 4; (vii) Percentage that this sound represents of the total instances of all sounds, (viii) Rank. It is sortable on the Sound and Total Instances fields.

**Letters Count.** This data table lists the frequency of occurrence of each of the 26 English letters. Its nine fields parallel those in Sounds Count. Notice that unlike the Spellings Count datatable this one deals only with the 26 individual letters. It is sortable on the Letter and Total Instances fields.

**Spellings Count.** This data table lists the frequency of the 154 different spellings from the Correspondences: Spelling to Sound data table. The large number of spellings arises primarily from three historical processes: First, the simplification of earlier clusters of vowel and consonant sounds in which the vowel and consonant letters are still written, as in the <mb> spelling at the end of words like *bomb* or the <io> spelling of schwa in words like *nation*. Second, the formation of doublets via twinning, assimilation, and simple addition. And third, the complicating effects of [r] on preceding vowels. This table is sortable on the Spelling and Instances fields.