Introduction to the Lexis Database
2016 Upgrade

> At the very least we need a systematic catalog of the elements of American English. Bound bases in particular remain undescribed in any useful way. We need a systematic catalog of them, just as we need a thorough catalog of the functioning sets of coelements in the language. It would be interesting and useful to have full explications of a large sample of the American English lexicon.
>
> *American English Spelling*, p. 462.

An Overview. The Lexis database attempts to address the needs outlined above, especially the final one. It is essentially a lengthy exercise in explication – that is, the analysis of written words into (i) written parts that contribute semiotic, morphological, or syntactic sense, (ii) any particles and vestiges, (iii) any historical processes such as assimilation that have affected their spelling, and (iv) the procedures that spellers must follow in spelling them. The following explication of *sufficiently* can illustrate: [su/b+f1+fic/e1+ient]+ly]1. This explication analyzes *sufficiently* into the prefix *sub-*, changed to the spelling <*suf*> via the historical process of assimilation with the letter <b> – and the sound [b] – being deleted and replaced with the particle <f1> and the sound [f], followed by the base +*fice1*. The number *1* following the bound base +*fice* indicates that this base is the first of at least two homographic bases spelled <fice>. The virgule indicates that when the expanded suffix *-ient* was concatenated to *suffice*, it did so via the procedure of silent final <e> deletion, while the adjective-to-adverb suffix *-ly1* was added to *sufficient* via the procedure of simple addition.

The Five Lexis Data Tables. The Lexis database contains five data tables. The first, Words, contains the lexicon of 129,042 words, each with its explication, as illustrated above with *sufficiently*. The second table, Bases, contains the 18,093 free and bound bases contained in those explications. The third table, Prefixes, contains the 252 prefixes from the explications in Words, and the fourth, Suffixes, contains the 1,168 suffixes. The fifth table, Particles, contains the 29 particles identified in Words. More detailed descriptions of the tables are given below.

Using Lexis. One typical use of Lexis would be to find a word's explication in the Words table and then, shifting to the other tables, finding more about its constituent elements, particles, processes, and procedures. If, on the other hand, you are interested not in elements but in simple letter strings, you could, for instance, search for the string <mpt> in the Word field, which would return 145 words, from *ademption* to *unkempt*. If you were interested only in words in which <mpt> is in a single element, you could search for the string <mpt> in the Explication field, which would return 116 words from *ademption* to *transumption*.

In the Comment fields in the bases, suffixes, prefixes, and particles tables there are key words that can provide some informative searches and that are listed below for the

separate tables. The prefixes, bases, suffixes, and particles tables can be sorted on the Instances field to isolate low and high frequency forms. The following discussion includes other suggested uses for the five tables.

The Words table contains only two fields: Word and Explication. One special point: Many of the words in Lexis have homographs, of which Words includes only a few. A second point: Lexis contains all the words from the index of words in my *American English Spelling,* which means that it contains an unusual number of words with spellings that are odd or peripheral to our English spelling system, such as *ngaiol*, a word borrowed from Maori referring to a type of small New Zealand tree.

The following symbols are used in the Explications field: Plus marks indicate internal boundaries between elements, vestiges, and particles. Virgules indicate that the following letter is to be deleted. A left square bracket marks the beginning of a prefix; a right square bracket marks the end of a suffix. Numbers following elements and particles distinguish homographic forms. For more about the process of explication, see "On Explication" in the Short Articles venue of this website.

The Bases Table contains the following seven fields: (i) Bases, (ii) Examples, (iii) Instances, (iv) Free, (v) Sense Links, (vi) Comments, and (vii) Relatives.

The Base and Examples fields are pretty much self-explanatory.

The Instances field gives the number of words in Lexis in which the base in question appears. A caveat: As changes were made in the various tables, the counts changed, and it's possible that some of the listed counts are off a bit. They are accurate enough for discussions of general patterns and relative frequency, but if you need precise figures, I suggest filtering the Words table on the base in question to double-check the count.

The Free field tags free bases. Untagged bases are bound.

The Sense Links field summarizes various senses carried by that base throughout its history, often running back to its proposed Proto-Indo-European (PIE) root, It does not attempt systematically to present the evolution of those senses. Rather, it tries to show some of the senses that have been associated over the centuries with words containing the base. Thus, it suggests how various metaphoric and metonymic relationships have produced the various senses. (For more on the roles of metaphor and metonymy in the evolution of our orthography, see the article "Orthography as an Evolving Complex System" in the Short Articles venue of this website.)

Searching for a semicolon in the Sense Links field returns all bases assumed to have evolved from a PIE root. The senses of the roots are given to the left of the semicolons. These assumed senses present a major complication: Primitive languages tend to be concrete and specific affairs, with abstraction and generality evolving over time. But in the comparative method used to reconstruct the senses of PIE roots, the need to find common themes that link cognate words from often widespread languages leads to proposed senses for roots that are often much more vague and diffuse, much more general and abstract than the senses probably were in PIE. In spite of any ahistoric generality and abstraction, the assumed PIE senses can help uncover the metaphoric and metonymic links that concern us.

The Comments field contains brief notes on that base's structure, such as the contraction and expansion of earlier forms. In Comments the keyword *Imitative* covers a multitude of types. Sometimes it's a straightforward imitation of a natural sound, such as in *moo* and *caw,* or *oh* and *ooh.* Sometimes it is not always clear what is being imitated, as with *jab* and *jam*, where it is perhaps more like sound symbolism or phonaesthesia. Other keywords in the Comments field: *Reduplicative, Folk ety*(mology)*, Alters, Varies, Nonterm*(inative) and *Term*(inative) *coform, Eponym, Merges, Redivides, ooo* (Of Obscure Origin), *Past* (tense), *Past participle, Archaic, Coined by, Converts, Trademark, Plural.* The keywords *Contracts* and *Expands* mean simply that the base in question is of a length different from another; whether it subtracts from a longer base or adds to a shorter one is not always clear. The Comments field also lists the source languages *Spanish, French, Anglo-Norman, British, Scots, Dutch, Arabic, Chinese, Hindi, Swedish, Norwegian, Finnish, Italian, Portuguese, Latin, Greek.* The Comments field also contains a few warnings about unusual deletions, tagged with the keyword *Check* – as in "Check +vi/e & +v/i/e" (at *vie*), necessary to catch the inflected forms *vied* and *vying*.

The English lexicon and its morphology are evolving complex systems, so an important part of explication is the attempt to capture some sense of the direction that things have taken and are taking. Much of this attempt can be seen by filtering on the keywords *Expands*, *Contracts,* and *Alters* in the Comment field. In the lists that are returned will be many cases in which historical bases have been expanded by the accretion of vestiges from earlier stems or contracted via the metonymic part-for-whole relationship, especially in words from the scientific-technical register.**The Relatives field** lists etymological relatives of the base in question. It is an exercise in *lexical cladistics* – the grouping of entities that descend from a common ancestor. But in a few lexical clades not all bases actually share a common ancestry, are not true cognates because over the centuries some non-cognate bases have been treated by human users **as if** they were cognates – due to analogy, folk etymology and similar drivers of language change, including simple error.

N.B. I have gone through the Relatives field three or four times, every time finding omissions, inconsistencies, and just plain errors. I assume that there are still errata lurking there. For instance, filtering on Relatives Is Null returns 3,378 bases. The vast majority of these I assume are singletons, with no relatives. But I'm sure that some of them do have relatives that should be tracked down – which I have not done.

Assumed PIE roots are listed in the Relatives field preceded by an asterisk – for example, **skep.* A few of these roots, as noted, are actually Latin, Greek, or Germanic roots, not Proto-Indo-European. The great majority of PIE roots are drawn from Calvert Watkins' *The American Heritage Dictionary of Indo-European Roots* (Boston: Houghton Mifflin, 2nd edn. 2000), listed with no additional tag. Some that are in Watkins' 1st edition but not the 2nd are tagged "(Watkins 1985)". Some that are in his 3rd edition but not in the 1st or 2nd are tagged "(Watkins 2011)". Several are drawn from *The Barnhart Dictionary of Etymology* (H. H. Wilson, 1988), tagged with the Pokorny page number – for example, "(Pok 345)". A few are drawn from Partridge's *Origins* and tagged "(Partridge)", and fewer, tagged "(Leiden)", are drawn from the interactive version of Julius Pokorny's seminal *Indogermanisches Etymologisches Worterbook* (Bern, 1959), available at Leiden University's website *http://www.indo-european.nl/*. Although I found out about it too late to use it in compiling the Bases field, another very useful on-line source for Indo-European roots and some of their reflexes is the Indo-European Lexicon available from the University of Texas at Austin at http://www.utexas.edu/cola/centers/lrc/ielex/PokornyMaster-X.html.

The PIE roots are all reconstructions, arrived at by comparing words in the various daughter languages and applying the rules of sound change that grammarians have developed – a remarkable accomplishment that is till on-going. No direct traces of PIE exist. All of the roots are assumptions. In his *American Heritage Dictionary of Indo-European Roots* Calvert Watkins uses the word *perhaps* often, as is often reflected in the Bases data table's many question marks. (For anyone looking for open questions to work on, I'd say *cherchez l' pointe d'interrogation*!)

As in any form of archaeology, etymological conclusions are often based on little hard evidence, and conclusions and assumptions can change drastically with the discovery of new evidence or the creation of new lines of thought. Thus, there can be, and is, considerable disagreement, not only about whether a given modern base descends from a certain PIE root, but also about the form and semiotic content of that root (and at times its very existence). In compiling the list of roots for the Relatives field, I've taken what might be called a "loose constructionist" approach – that is, since I am more interested in finding plausible unifying links than in determining true etymons, I have assumed that if one respectable source finds a certain link plausible, even though others do not, I tend to include it, usually with a question mark.

In the Relatives field, Related elements are marked with leading plus signs, which can be thought of as substitutes for italics. Due to font limitations, some special characters have substitute symbols: Schwa is represented with @; long vowels are represented with capital letters.

The treatment does not pretend to be exhaustive, whatever *exhaustive* might mean here, since in some cosmic sense and along various dimensions all lexical elements are related to all others. Of course, all bases with the same PIE root are assumed to be related, so bases returned by a search in the Relatives field for a specific root are assumed to be related to one another to some degree. In general the more similar are their Senses fields, which deal with semiotic content, the more closely related are the bases. In an even more general sense, a principle of transitivity applies here: If *a* is related to *b*, and *b* is related to *c*, then *a* is related to *c*. To pursue these relationships further you can consult the family trees provided by Watkins, either in his *The American Heritage Dictionary of Indo-European Roots* or in his Indo-European appendices to the *American Heritage Dictionary*. Relevant information can also be found in the Indo-European Lineages venue of dwcummings.com or the University of Texas Indo-European Lexicon mentioned earlier.

The Relatives field usually lists only bases, not affixes. When prefixes are listed – for instance, at *pert1* there is "+[a4", the prefix [*a4-* in the Prefixes table – it means that there is at least one word containing that base that is an aphetic form of an earlier word with that prefix, as *pert*1 itself, which is from ME *apert* [a4+pert. To find other affixes that have been converted to bases, you can filter the Comment field on *Convert*.

Acknowledgements. This is very much a derivative study, based on the primary research of others: James Murray, Henry Bradley, Calvert Watkins, Julius Pokorny, Eric Partridge, and the editors of *The Barnhart Dictionary of Etymology*. Although the explications are my own work, much of the Bases table, especially the Sense Links and much of the Relatives fields, gathers together findings from many earlier studies.

The Prefixes table contains the 252 prefixes found in the explication of the 129,042 words in the Words data table. It is a rather long and eclectic list that includes, among other things, some prefixes from very rare adoptions – such as the plural markers *ema-* (*emalangeni*) and *ma-* (*makota*) from Africa, and the noun markers *mi-* (*mikado*) and *sa1-* (*samurai*) from Japanese. Such prefixes are obviously quite peripheral to the English affix system, but there they are in their adopted words. Since the primary motive of explication is to highlight potential unifying links in the English lexicon, it seems best to explicate to these alien prefixes, rare and exotic though they may be.

In addition to the Prefix field, the Prefixes table includes Examples, Instances, and Comment fields. In the Comment field, since many prefixes carry semiotic content, their senses are given, in quotation marks: [*ultra* "Beyond, beyond the norm." In the

Relatives field the listing of relatives is pretty much limited to clear cut assimilations and variations. The Comment field contains a number of keywords: *Contracts, Alters, Expands, Varies,* and *Marks* and *Forms*, which show syntactic function. An especially important keyword is *Assimilation*, which tags assimilated forms of several prefixes. The Comment field also contains the keyword *Search,* followed by the syntax for filtering to given assimilated forms – for example, at [*ac*1-, you find "Search [a/d+c+".

The treatment in Prefixes leads to clades that are small and of fairly recent formation. However, a number of prefixes descend from Indo-European roots, identified in the Comment field with a leading asterisk – for instance, **en.* Since prefixes descending from the same PIE root can be assumed to be related, you can find older, larger clades going back to PIE by filtering the Comment field on a specific PIE root – for instance, "Like "*[*]en*".

Finally, the Comment field lists a number of source languages and country names: *Old English, German, Latin, French, Spanish, Greek, Italian, Romance, Russian, Japanese* – and from Africa: *African, Swaziland, Bemba, Bantu, Zaire, Lesotho*.

Definition of Prefixes. In his *English Word Formation* Hans Marchand uses a rather restrictive definition of *prefix* ("Prefixes are bound morphemes which are preposed to free morphemes" [129]) and discusses only 65 (129-208). In his *Origins* Eric Partridge uses Webster's similar definition of *prefix*: "One or more letters or sylables combined or united with the beginning of a word to modify its signification, as *pre-* in *prefix, con-* in *conjure*" (821). Partridge goes on to say that "Strictly, a prefix should consist of either a preposition or adverb", and he omits what he calls the "false prefixes of science", which he describes as not prefixes, but abbreviations. Still, he lists more than 340 prefixes, including many homographic and variant forms – for example, he lists seventeen different prefixes spelled <a>.

So the question of what a prefix is remains somewhat open. 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 *electro+* as an example of a combining form (at "combining form"), in the main word list it is labeled "prefix," as apparently are all other combining forms. However, the *RHUD* and *W3* both distinguish carefully between prefixes and combining forms. At a different extreme the editors of *Prefixes: and Other Word-Initial Elements of English* (Laurance Urdang and Alexander Humez, Old Lyme, CT:Verbatim Books, 1998) 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 confusion: 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. For the sake of simplicity, I treat all numerical elements as combining forms – that is, bases, usually bound – and restrict prefixes essentially to prepositions (*in2-, ad-*), negatives (*in1-, non-, un1-*), adverbs (*se-, per1-*), a few derivationals, (*en1-* and *be-*), and even fewer inflectionals, such as those in

samurai and *mikado*.

There can be indecision whether the Romance euphonic <e>, as in *escalate*, is an initial particle or a prefix. Arguing against the former is the fact that all other particles are medial, not initial. Arguing against the latter is the fact that euphonic <e> does not have any semiotic or syntactic content. For now I explicate the euphonic <e> as a particle, listed as *e4* in the Particles table.

Converted Prefixes. The practice of synecdoche and conversion in English has led to several cases in which one-time prefixes have been converted to bases. Some examples: The base *com5* contracts *communication* (as in *intercom*), and *com7* contracts *commissioned* (as in *noncom).* Both convert the prefix [*com-* to a base. In *abo* ab2+o]3 the base *ab2* is clipped from *aborigine* [ab1+orig+in]02+e]3. *Selsyn* sel4+syn1 contracts "sel(f) + syn(chronous)" [syn+chron+ous], converting the prefix [*syn*. In *cistron* cistr+on]08 the base *cistr* contracts the phrase "cis-trans test". Since [*cis-* and [*trans-* are both prefixes, the bound base *cistr* is another example of conversion. The very common free base *pro1* converts the prefix *pro1-* in *professional*.

Often the prefix merges with part or all of the following base in the conversion: In *propane* prop2+ane]3 the base *prop2* contracts *propionic* [pro2+pion2+ic]1. In *praetor* *prae+tor*], *pretor* pre+tor] and their derivatives, the *prae+* and *pre+* convert the Latin prefix *pre-* to a base and absorb the <e> from a now lost Latin base meaning "go". *Combo* comb3+o]3 contracts *combination,* with the prefix *com-.* Similar examples: comfy, *commie*, *comp3, condo*. *Pregnant* pregn+ant]1 is ultimately from a Latin source that contains the prefix [*prae-* plus the source of our modern base *gn2* "born". Related are slang terms such as *preggers, preggy,* and *preggo.*

In the other direction, bases are sometimes converted to prefixes, as is apparently the case in *sovereign*, which I explicate as [sove+reign. The word has evolved from ME *soverain,* influenced by *reign* . The ME *soverain* is from OF *soverain*, from Vulgar Latin **superanus*, which would be the Latin *super* "above" plus the adjective suffix *-anus*. *Sovereign* is a good example of folk etymology: As the final syllable was respelled <reign>, the respelled Romance suffix *-ain* was converted from suffix to base. It would also follow that the first syllable, which started out as a base has converted to a prefix, a variant of the English prefix *super-.*

The explication of prefixes that descend from the Latin preposition *ad* raises some questions. At one extreme are those words in which the Latin preposition had already become used as an assimilated prefix in Latin, words like *accent* and *accident.* At the other extreme are words from French phrases with the contracted French preposition *a*, with no double consonant: *abase, abate, abut*. The former I explicate as [a/d+; the latter as [a3+. Problems arise with words between these extremes, with words like *acclimate*, which developed the <cc> in French, and *accompany*, which developed it in English. In such words there never was an assimilation of *ad-.* But rather than positing a separate prefix *ac-,* I recognize the power of analogy and folk etymology and treat

such words as if they were assimilated forms of *ad-*. Basically, if there is a double consonant, or its equivalent, I'm inclined to explicate to [a/d+-, motivated by analogy. If there is no double consonant, I'm inclined to explicate to the closely related [a3+.

The Suffixes Table contains the 1,168 suffixes explicated out in Words. The relationships among suffixes can be quite complex and uncertain. The treatment of suffixes in the Relatives field is probably best thought of as "notes toward," consisting of incomplete work in the Relatives field and possible relatives, tagged in various ways in the Comment field : *coform, cf, varies, expands, contracts, alters*, etc. Also further good candidates for expansion are tagged *vestige*.

In the Comment field there are a number of potentially useful search and filter strings. The following are the more common for the country of origin: *Greek*, *Latin*, *GrecoRom* (which equals GrecoRomance, which is Greek, Latin, and modern Romance languages), *Rom*(ance), *Italian*, *Spanish*, *Portuguese*, *Germanic*, *German*, *English*, *British*, *OE* (Old English), *Scandinavian*, *Russian*, *Slavic*, *Hebrew*, *Yiddish*, *Arabic*, *Semitic*, *Hindi*, *Japanese*.

In what might be called a "processes" group in the Comment field there are the following: *Varies, Marks, Vestige*, *Term*(inative) and *Nonterm*(inative), *Coform*, *Converges, Combines*. So far as the tags *Contracts* and *Expands* are concerned, the same caveat stands here as with the Bases table: Whether we're dealling with a contraction or an expansion is not always clear.

The Comment field also lists several strings dealing with grammar: *noun*, *verb*, *adj*(ective), *past* (tense), *pres*(ent participle) , *part*(iciple), and many others. Related are *agent*, *instrument*, *frequentative*, *diminutive*, *augmentative*, *comparative*, *superlative*, *pejorative*, *plural*, etc.

In the register group are *chemical*, *scientific*, *technical*, *jocular*, *familiar*, *informal*, *intimate*.

I tend to explicate so as to recover as much as possible of potentially motivating material from the words. One result of this is that sometimes explicating to a base leads to suffixes that are quite exotic and peripheral, such as the following from Hebrew: -o]7*, -oh*], -os]3*, -ot*]6*, -oth*], -u]5.

Nonterminative Suffixes. The notion of nonterminative suffixes may at first seem odd. They tend to occur as nonterminative coforms in sets like {-abil]+, -able]} and {-os2]+, -ous]}: *availability, available*; *generosity, generous.*

The Particles Table lists the 29 particles identified in Words and contains the following fields: Particle, Examples, Instances, Comment. The Comment field contains filterable tags that add descriptive information: *linking, via twinning, via assimilation, via*

<i> to <y>, via <y> to <i>, euphonious, varies, native. For more on particles see “On Explication” in the Short Articles venue of this website.

*** * * * * ***

Available Text Files. The on-line version of the database has limited filtering and sorting capabilities. If you need to do more elaborate filtering and sorting – using Boolean operators, for instance – feel free to download the text files of the tables and load them into your database or spreadsheet program.

The only requirement is that any public use of the Lexis database be publicly acknowledged and documented.

A Work in Progress. The Lexis database can only be seen as very tentative. My hope is that later work will lead to more formal and systematic ways of answering the question this work raises. Changes and corrections are expected and encouraged, and the word *tentative* in the first sentence of this paragraph is carefully and deliberately chosen

In a project of this size, one that has stretched over so many years, a major problem is maintaining consistency of analysis. If in using Lexis you discover inconsistencies of any kind, please point them out to donwcummings@charter.net. Beyond that, there is an immense potential for all kinds of errors, from pesky typos to just plain mistaken explications. I would appreciate hearing about any of those you might discover. As inconsistencies and errors are caught and corrected, we plan to update the tables on the site. I would also appreciate any suggestions or comments.

D. W. Cummings

donwcummings@charter.net