

Development of the Word Dissector Application

Sources and Initial Development

The principal initial source for the word list and morpheme resources used to develop the Iowa Reading Research Center (IRRC) Word Dissector was the academic-word database developed by Lane et al. (2019). Lane et al. compiled 6,597 unique K-12 academic words from eight sources, including the Academic Word List and widely used word lists for English language arts, mathematics, science, and social studies. Their study and accompanying morpheme list organized high-frequency prefixes, derivational suffixes, Latin roots, and Greek roots and provided primary spellings, alternate spellings, meanings, and example words. The database also contained fields for inflectional suffixes. For the specific purpose of calculating morpheme frequency, however, Lane et al. did not count inflectional suffixes and removed most redundant inflected word forms before conducting the frequency analysis.

Using these materials as a starting point, the research team at the IRRC collaboratively developed the Word Dissector entries. For each selected word, the team entered the component morphemes and the meaning associated with each component; they then added information about morpheme type and language of origin. Lane et al.'s analyses and morpheme lists served as the initial instructional reference. The Oxford English Dictionary (OED) served as the primary source for verifying and revising word-specific etymology and formation, and the morpheme list in Stockwell and Minkova (2001) served as a supplementary reference. The project dataset included inflectional endings, which were retained in the finalized Word Dissector.

Review, Evidence Hierarchy, and Stopping Rule

The IRRC research team reviewed the entries word by word. During this review, the team considered where to place morpheme boundaries, which spelling of each morpheme appeared in the target word, the meaning and language of origin of each element, the classification of each element as a prefix, root, or suffix, and the consistency of the analysis across related words. When the sources differed, the OED's word-specific account generally guided decisions about formation, classification, origin, and context-specific meaning, while the Lane-based materials and Stockwell and Minkova (2001) informed broader morpheme-family relationships and instructional glosses.

The team first identified the immediate base or bases involved in the formation of the target word. When an OED entry did not provide a full etymology, the team followed its cross-references to relevant prefix, suffix, base-word, and related etymological entries. An immediate lexical base was retained unless a linked entry supported a further division into recurring, meaning-bearing elements that were visible in, or systematically related to, the target spelling. Thus, *nonunion* is represented as *non* + *union* and *recommence* as *re* + *commence*, whereas the linked entries for *reimbursement* support the finer analysis *re* + *im* + *burse* + *ment*. This procedure allowed the team to distinguish a warranted deeper analysis from a division based only on a coincidental sequence of letters.

When a source word, or etymon, could itself be analyzed into smaller components, the team followed the linked entries only to the deepest instructionally relevant level. A further division was included when it exposed a recurring element that contributed identifiable meaning and could be represented consistently across a word family. A further division was not included when it reflected only the internal history of a remote etymon, introduced material no longer recoverable in the target word, or required treating nonmeaningful residual letters as morphemes. The displayed components were also checked against the target spelling: they were expected to account for the complete form without duplicating letters or leaving unexplained material. Apparent errors or inconsistencies in the source materials were documented and shared with the project team for resolution.

For example, the OED traces *hemoglobin* to the earlier form *haemato-globulin*, which it identifies as a variant of *haematio-globulin*. The linked entries relate *haematio-globulin* to *haematin* and *globulin*. Additional entries analyze *haematin* as Greek *haimat-* “blood” combined with *-in* and *globulin* as *globule* combined with *-in*. These linked entries establish the semantic and etymological relationships represented in the Word Dissector. For instructional display, however, *hemoglobin* is analyzed as *hem(o) + glob + in*. This representation preserves recognizable elements in the target spelling while identifying *hem(o)-* as the element associated with “blood,” *glob-* as the element associated with a “round mass” or “ball,” and *-in* as a noun-forming element used in names of substances.

The analysis of *hem(o) + glob + in* is not a complete transcription of every historical stage reported by the OED. The OED traces *globule* further to Latin *globulus*, historically analyzable as *globus* combined with the diminutive suffix *-ulus*, reflected in English *-ule*. The Word Dissector does not display *glob + ule + in* because *-ule* is no longer separately recoverable in the shortened form *hemoglobin*. Similarly, the *-in* contained in the historical source *haematin* is not represented as an independent element within *hem(o)-*. These elements belong to the internal formation of more remote etymons rather than to the instructionally useful decomposition of the target word itself.

More generally, some words entered English as already-formed lexical units even though their earlier etymons contain smaller Greek, Latin, French, Germanic, or other historical elements. Separating those elements can produce a finer-grained analysis than would be used in a strictly synchronic description of present-day English. This was an intentional instructional decision. The breakdowns are intended to make recurring formative elements visible so that learners can examine how words were formed, recognize relationships across word families, and use those relationships to reason about meaning. They are not intended to imply that every word was assembled from the displayed elements in modern English or that contemporary readers necessarily process every displayed boundary.

Representation of Surface Forms, Allomorphs, and Word Families

Consistent with this instructional purpose, the Word Dissector preserves the spelling that appears in the target word whenever possible rather than replacing every spelling variant with a single canonical form. For example, *cap*, *cept*, *ceive*, *cip*, and *cup* may be treated as members of the same morpheme family because they share an etymological source and related meanings, while each entry displays the spelling found in the target word. Related spelling variants, or allomorphs, remain linked through their etymological and semantic

relationships without erasing their orthographic differences.

Parentheses are used selectively to mark variable orthographic material within a morpheme family. Depending on the target word, the parenthesized letters may be present or absent; they are not treated as a separate morpheme. Thus, forms such as *lat(e)*, *stat(e)*, *hem(o)*, and *meteor(o)* preserve relationships across a family while accounting for the spelling of the individual target. Sequences that merely resemble a familiar affix are not automatically coded as affixes. For example, *translate* is represented as *trans* + *lat(e)* and *substrate* as *sub* + *strat(e)* rather than assigning an independent causative *-ate* where the letters belong to the borrowed stem. Likewise, a larger base is retained when a proposed remainder is not meaning-bearing, as in *principle* rather than *princip* + *le*.

Systematic spelling alternations across related forms are represented by the form that actually occurs in each word. In the *unify* family, for example, *-fy*, *-fi*, and *-fic* are linked as members of one derivational family: *unify* is represented as *uni* + *fy*, *unified* and *unifies* as *uni* + *fi* + *ed/es*, *unifying* as *uni* + *fy* + *ing*, and *unification* as *uni* + *fic* + *ation*. Comparable alternations may occur before another suffix, as in *conduct* + *ive* but *conduct* + *iv* + *ity* in *conductivity* without further dividing the base *conduct*. Such surface forms are not treated as unrelated morphemes merely because their spellings differ.

The analysis of related words was made as consistent as the evidence allowed, but family consistency did not override word-specific formation. The same letters can belong to different boundaries in different words, and a familiar suffix was not added if doing so duplicated material already contained in the root or stem. For example, the shared stem in *legislation*, *legislative*, *legislator*, and *legislature* is represented by *legis* + *lat-*; accordingly, *legislation* is *legis* + *lat* + *ion*, not *legis* + *lat* + *ation*.

Coding Conventions for Derivational and Inflectional Patterns

Following Lane et al. (2019), *-tion* and *-sion* are treated as surface variants of the *-ion* family, while *-ation* is retained as a separate instructional suffix. Although *-ation* historically reflects a relationship between *-ate* and *-ion*, it is commonly recognized as a suffix in English, including in words that do not have a corresponding form in *-ate*. In the Word Breakdown tab, *-ation*, *-ion*, *-tion*, and *-sion* are coded as noun-forming derivational suffixes, and Latin and French are recorded as relevant source histories. The displayed boundary is determined for each target word. When final *-at-* belongs to the preceding lexical root, that stem is followed by *-ion*, as in *stat* + *ion*, *automat* + *ion*, and *legis* + *lat* + *ion*.

A similar distinction is made in the *-logy* family. Following Lane et al. (2019), *-logy* is retained as an instructional suffix in nouns in which the surface sequence carries the meaning “science or study of,” as in *biology*. Related derivatives display the form that occurs before another suffix, such as *bio* + *log* + *ic* + *al* in *biological* and *bio* + *log* + *ist* in *biologist*. This convention preserves the visible form while maintaining the relationship among members of the same word family.

Homographic endings are coded according to their word-specific formation rather than by spelling alone. For example, the noun-forming *-y* in *philosophy* and *policy* is associated with French *-ie*, while other occurrences of final *-y* may be native adjective-forming suffixes or part of a larger suffix such as *-logy*. In plural forms ending in *-ies*, the Word Dissector

displays the surface plural form without also retaining a separate final *-y*. *Philosophies* is represented as *philo + soph + ies*, not *philo + soph + y + ies*. Similarly, *chemistry* and *componentry* contain *-ry*, whereas *contrary* is analyzed with *-ary*; visually similar endings are not assumed to be the same suffix.

The finalized Word Dissector retains inflectional endings that occur in the target word. Plural *-s*, *-es*, and *-ies* are glossed as “forming plural nouns.” Verbal *-s* and *-es* are glossed as “forming third-person singular present-tense verbs.” Inflectional *-ed* is glossed as “forming past-tense and past-participle forms of regular verbs,” and *-ing* as “forming present-participle or gerund forms.” These spellings account for the complete target form but do not imply that each surface variant constitutes a separate underlying inflectional morpheme family. Their inclusion is an instructional and representational decision distinct from Lane et al.’s decision not to count inflectional suffixes in the morpheme-frequency analyses.

Classification of Morphemes and Lexical Bases

The Word Dissector uses three broad morpheme categories: prefix, root, and suffix. When word-specific evidence identifies an element as a prefix or suffix in the target formation, that classification is retained. The OED also frequently describes bound Greek and Latin elements as combining forms. Because “combining form” is not a separate category in the finalized Word Dissector, an element described only as a meaning-bearing combining form is coded operationally as a root, in alignment with the terminology used in the “Effective Morphology Instruction for Secondary Students Across Content Areas” eLearning module. Thus, *uni-* in *unify*, *unity*, and *uniform* and *auto-* in *autobiography* are coded as roots, whereas *syl-* in *syllable* is retained as a prefix because the OED identifies its Greek source *syn-* as a prefix in that formation.

Free lexical elements that function as bases in compounds and larger borrowed lexical bases are likewise coded operationally as roots when no separate free-base category is available. Examples include *union* in *nonunion*, *commence* in *recommence*, *engine* in *engineer*, and a retained whole base such as *quadrant* when a smaller English suffixal analysis is not supported. This convention does not claim that free bases, borrowed stems, combining forms, and roots are identical in every linguistic framework. It provides the closest available project category for elements that carry the central lexical meaning.

Entries are not required to contain exactly one root. A word may contain more than one meaning-bearing root or combining form, as in *autobiography*, and a well-supported analysis may contain no element coded as a root if the word is adequately represented by a prefix and suffix. Conversely, the team did not create an artificial root solely to satisfy a one-root requirement. The number and location of roots were determined by the evidence and the project’s operational categories, not by a fixed template.

The team distinguished whole-word formation from morpheme-level origin. A word formed within English can contain elements of Greek, Latin, French, Germanic, or other origin. Identical spellings were not assumed to represent the same morpheme. For example, *in-* may be a Latin prefix in one word, whereas *in* in the compound *income* is an adverbial element of Germanic origin. Likewise, visually identical forms such as *anti-* or *pro-* can represent different historical elements and meanings in different words. Classification, meaning, and origin were therefore determined from the word-specific formation rather than from orthography alone.

Language-of-origin labels are assigned to the displayed element, not automatically to the whole word. When more than one relevant source language is recorded, the labels identify source histories relevant to that element; their order should not by itself be interpreted as a complete chronological route of transmission. If a larger lexical base is retained, the origin label reflects the source history of that retained base rather than every more remote ancestor that could be reconstructed.

Assignment of Morpheme Meanings

Morpheme meanings in the Word Breakdown tab are assigned at the level of the target word. When the OED provides sufficient evidence that one sense within a broader morpheme family is active in a particular word, the IRRC research team uses that context-specific gloss. For example, the broad source gloss for Latin *dis-/di-/dif-* includes “not,” “apart,” and “absence of,” whereas *di-* in *dimension* is glossed as “out” because the relevant Latin verb means “to measure out.” Similarly, Lane et al. group *con-/co-/col-/com-/cor-* under the broad meaning “with, joint,” whereas the OED identifies *con-* in *constitute* as intensive. Occurrences of *con-* in the *constitute-constitution* family are therefore glossed as “intensive; strengthens the meaning of the base,” while other members of the broader family retain a different gloss unless word-specific evidence supports the intensive sense.

A morpheme gloss describes the contribution of the displayed element rather than serving as a complete dictionary definition of the word. When the evidence does not support a narrower interpretation, the broader source gloss is retained. Different contextual glosses do not by themselves indicate different morpheme families; family membership is based on etymological relationship, recurrent form, and shared semantic history.

Part-of-Speech and Sense Selection

When a target form can function as more than one part of speech or has more than one relevant sense, the IRRC research team considers the part of speech and sense indicated in the source entry when available, the OED’s word-specific accounts, the surrounding word family, and the instructional role of the entry. If the same morphological analysis applies across the relevant uses, one breakdown is retained, and a suffix gloss may identify more than one possible outcome category, such as adjective and noun. The OED’s listing of multiple parts of speech does not by itself require duplicate Word Breakdown entries.

When the grammatical status of an ending changes with the intended part of speech or sense, the intended use controls the coding. For example, *-s* or *-es* is glossed as a plural ending in a noun entry but as a third-person singular present-tense ending in a verb entry. Likewise, *principled* in its common adjectival sense “guided by principles” is analyzed as the noun base *principi(e)* plus adjective-forming derivational *-ed*, glossed “having or characterized by X.” If the same surface form were analyzed as the past-tense or past-participle form of the rare verb *principle*, *-ed* would instead be inflectional. Thus, the part-of-speech label is considered together with the base and the function of the ending rather than used as an automatic rule.

When the available source information does not identify a single intended use and competing analyses cannot be combined, the team selects the analysis that is best supported by the OED, most consistent with the relevant word family, and most useful for the instructional purpose of the resource. The alternative analysis and the reason for the

decision are documented in project notes when the ambiguity is consequential. The final Word Breakdown presents one purposeful instructional analysis for each entry rather than displaying multiple unresolved alternatives.

Relationship Between the Morpheme Cards and Word Breakdown Tabs

The Morpheme Cards tab is retained as a source-reference table compiled during the resource's initial development. It preserves the terminology, classifications, broad family-level meanings, examples, and language-of-origin labels used in the source materials and working files, including the category combining form. These source-reference labels should not be interpreted as the final classification or gloss assigned to every occurrence in the Word Breakdown tab.

The Word Breakdown tab contains the final word-level instructional analysis. Its spelling, boundary, classification, meaning, and origin fields may therefore differ from a Morpheme Cards entry when the OED provides more specific evidence or when the project's final coding conventions require a different operational category. For example, the source-reference entry for *-ation* may retain a Latin-only origin label, whereas the final Word Breakdown convention records Latin and French as relevant source histories. Preserving the source table allows users to trace the project's initial resources without requiring the final analyses to reproduce every source label unchanged.

Interpretive Status of the Entries

When the available evidence supported more than one plausible analysis, the IRRC research team reached a collective determination by considering the OED's account, the Lane-based materials, Stockwell and Minkova (2001), patterns across the relevant word family, the spelling displayed in the target word, the intended part of speech and sense, and the instructional usefulness of the alternatives. The IRRC research team collaboratively developed the resulting entries as research-informed instructional analyses. They are designed to support learners' examination of meaningful word relationships and may differ from segmentations developed primarily for historical-linguistic, synchronic, psycholinguistic, or computational purposes. A breakdown presented in the Word Dissector should therefore be understood as a purposeful instructional representation rather than the only possible linguistic analysis.

Resources

- Lane, H. B., Gutlohn, L., & van Dijk, W. (2019). Morpheme frequency in academic words: Identifying high-utility morphemes for instruction. *Literacy Research and Instruction*, 58(3), 184-209. <https://doi.org/10.1080/19388071.2019.1617375>
- Oxford English Dictionary. (n.d.). Oxford English Dictionary. Oxford University Press. <https://www.oed.com/>
- Stockwell, R. P., & Minkova, D. (2001). English words: History and structure. Cambridge University Press.