Beginnings of a Database of Ancient Greek Word Formation

A thesis submitted by

Ingrid Barnes

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Digital Tools for Premodern Studies

Department of Classical Studies

Tufts University

February 2023

Adviser: Gregory Crane

Abstract

This paper outlines some of the work done towards a database of Ancient Greek word formation, using the Logeion version of the Liddel Scott Jones (LSJ) dictionary as a basis. Tools such as Morpheus are available for lemmatizing inflected forms of Ancient Greek, but no such tool exists for derivational or compounding processes. A database is a step towards such a tool for word formation, and the data itself is of use both to students and scholars of Ancient Greek. A preliminary evaluation of the LSJ suggests that the breadth of the entries can be reduced significantly if describing them using their constituent parts.

Table of Contents

List of Tables

Introduction

The goal of this project is to create an initial, extensible database of Ancient Greek words and their morphemes, by systematically separating each Greek word into its constituent morphemes, and then connecting and grouping all words that share those morphemes. This has been systematically done for inflection,[1] but not for the word formation processes of derivation and compounding.

At its furthest extent, this would mean connecting all surviving Ancient Greek words to each other, detailing all Greek derivational morphemes and their relevant transformations, and relating all compounds to their constituent lexemes. In a narrower scope, I begin with lexical entries in the Liddle and Scott Jones Greek Dictionary (LSJ), which provides a solid base, and eventually additional lexical entries from other dictionaries may supplement this.

The main purpose of my work so far has been to lay the foundations for an annotated database of Ancient Greek words and demonstrate its worth, both in itself and its potential uses for analysis. Here, I lay out a case for such a database, similar efforts that have been made, the work I have done towards its completion (both successful and unsuccessful), and further work to be done.

[1] The Morpheus parser, for example, links inflected forms to their headwords, e.g. *πέμπεις "you send"* to *πέμπω "I send"*. See Crane (1991) and https://github.com/PerseusDL/morpheus/

Reasons for Doing This & Further Explanation

Digital projects, including corpora, have already demonstrated how digital projects can be useful for both scholars and students. The Cambridge Grammar of Classical Greek (CGCG) states that "digital search corpora such as the Thesaurus Linguae Graecae [TLG] and Perseus under PhiloLogic were of great help in finding suitable material."[2] Likewise, the reasons for a systematic analysis of Ancient Greek word formation processes are both pedagogical and scholarly–the analysis a particular text, author or corpus for the most frequent (or unusual) morphemes used, and attaining a better understanding of the word formation processes and lexicon of Ancient Greek, both morphologically and semantically. Current tools are lacking for these purposes.

"Pedagogical" Reasons

A database of word formation is useful for constructing frequency lists for Ancient Greek, either generally or for a specific author or text. However, rather than creating lists of Greek headwords, it would be possible to create a list of Greek morphemes. Headword frequency as a measure for pedagogical purposes is often insufficient. Many words, from common words to hapax legomena, are merely composed of other Greek roots and affixes that are common. Word frequency count will miss these, but a well refined morpheme frequency count will not. Knowledge of word formation processes, including the possible morphophonological changes, helps in extrapolating the meaning of known words to unseen words in sight-reading.

[2] van Emde Boas et al. (2019), p.xxxiii.

A focus on word formation also allows and encourages a better understanding of the Ancient Greek perspective, rather than simple translational knowledge. The connections between the words *φίλος* “friend”, *φιλία* “affectionate regard, friendship”, and *φιλέω* “love, regard with affection” is fairly clear even in translation, but the derivational processes are often unexplained in textbooks, and finding examples of the same pattern can be arduous since grammars are necessarily limited in their examples.

Another (admittedly more obscure) example is *μυοφόνον*. The definition in the LSJ is simply “wolfsbane”, even though it is more appropriately “mousekiller” to the Ancient Greeks (or perhaps “mousebane”), and *λυκοκτόνον* “wolfsbane” or “wolfkiller” also exists.[3] It is unclear how many words have opaque but explicable definitions. A database would better enable a discussion and study of how Ancient Greek words are semantically constructed in an organized, corpus based fashion.

“Scholarly” Reasons

There is already work in the study of word formation that demonstrates a use for such a database. There is a derivational morphology project for Latin - Latin Word Formation[4]. It is a MySQL relational database, including both derivation and compounding. The scope, in their own words, “...is to assign a WFR to each morphologically-complex lexeme (i.e. one word morphologically derived from another

[3] This is based on the LSJ entries. It is possible there are better identifications for what plants these words refer to.
[4] Litta & Passarotti (2018). See https://github.com/CIRCSE/WFL for the database itself, and https://wfl.marginalia.it/ for a presentation of it.

word) and to link each complex lexeme to its ancestor. All those lexemes that share a common (not derived) ancestor belong to the same 'word formation family' ."[5] I believe a similar project for ancient Greek is a natural extension of this idea.[6]

One of the inspirations for this project is *Ancient Greek Verb-Initial Compounds* by Olga Tribulato.[7] It contains a comprehensive discussion of verb initial compounds, including their history and semantics, and an appendix listing all of these compounds. Reading through this yielded two ideas: first, that this appendix could be a database of some kind; second, that a larger database could have been queried to yield this same list of compounds, and yet more lists if sufficiently annotated.

In the *CGCG*, there is a section about word formation, but the digital works it cites as being helpful for finding material, the TLG and Perseus under Philologic, are not made for querying this sort of material. There are etymological works that contain word formation information, such as *The Etymological Dictionary of Ancient Greek* or *La Formation des Noms de Grec Ancien*,[8] though these are dictionaries of text and not queryable digital works.

Having a database of word formation would also enable scholars to quickly access the data they are looking for. For derivational suffixes, one could simply look at the last characters of lemmatized words. However, extra work must be done to sort out which ones are derived or simply part of the stem. Suffixes that are part of a

[5] ibid.
[6] I must note that I did not encounter this project until midway through my own project, so it is of limited inspiration for my structure up to this point. Nevertheless, it is a useful template for a word formation project. It delves into theoretical issues and other problems that I have pushed down the road, and has a good corpus-based distillation of word formation rules.
[7] Tribulato 2015.
[8] Beekes 2011 and Chantraine 1933.

concatenation are obscured (such as *-ιζω* in *-ισμος*). A database essentially is having this sort of work pre-done so that others can cleanly access it.

A database would also help to corroborate claims about Greek word formation. For example, the *CGCG* makes the assertion “simplex verbs are only compounded with preverbs”, which is quite the absolute statement for a lexicon of over 100,000 words. Though I find no reason to doubt the statement, a database can provide the data to corroborate this assertion, as well as corroborate other statements like it, or perhaps even find them.

The Base Data - LSJ, Logeion and the Limitations of Dictionary Data

I use Logeion LSJ data[9] as the main basis for my work. The data is in Unicode, and the LSJ is wide-ranging, including many obscure and unused words such as those from the lexical work of Hesychios of Alexandria. The Logeion IDs also connect my annotations back to the Logeion data, so that they may be used for any other work that also uses them. I use Bailly 2020, an Ancient Greek-French dictionary, as supplementary data, but only matched to LSJ headwords.[10] Some entries don’t match LSJ entries perfectly, either because of differences in accentuation or chosen form,[11] or homonymic entries might be condensed into one or split into multiple. Bailly often has

[9] Logeion is data from the Perseus Project, which has subsequently been further edited and curated by Helma Dik. See https://github.com/helmadik/LSJLogeion for the data itself, and https://logeion.uchicago.edu for an interface.
[10] Bailly 2020 is a digitized form of Anatole Bailly’s *Dictionnaire grec-français*. For the releases, see https://github.com/latin-dict/Bailly2020/releases. The data is also incorporated into the Logeion interface.
[11] e.g. one uses the middle voice ending -ομαι, the other the active voice -ω, or one uses the singular and the other the plural.

morphological splits where the LSJ lacks them, as well as explicitly states the roots contained in derived words.

Dictionary entries typically rely on the reader's knowledge of the language rather than explicitly stating things about the entry. Parts of speech are implicit, and the gender of nouns is implied by the article. They also have some morphological analysis in many of their entries–usually only the latest part of the formation of the word, and also expect the reader to do the work in parsing possible ambiguities of the analysis. The sequence *αν-* at the beginning of such an analysis is representative of this ambiguity: it can either be the prefix *ανα-* before a non-alpha vowel, or it might be an alpha privative before a vowel. It is left to the reader to disambiguate which it is, and occasionally the entry contains definitions for both analyses.

The LSJ, and more frequently Bailly, will sometimes give the etymology of a compound word as multiple words. In particular, Bailly's etymological notes, with their frequency of inclusion and accuracy, may be sufficient in themselves for the basis of a database. However, it is not entirely linguistically accurate to represent compounds as the result of whole words being concatenated, as the Bailly entries might imply. The words can be of a different part of speech from the given constituents. To use an earlier example, the *ιππος* "horsed" in *ἄφιππος* "off-horsed" is not the noun *ἵππος* "horse",[12] but an adjective derived from it. Bailly gives it related to *ἀπό*, *ἵππος*, but to directly link these in a database would be misleading without specifying the derivational process. There

[12] Definitions here are meant to be illustrative of the forms, rather than true to the LSJ definition. The LSJ gives it *ἄφιππος* as "unsuited for cavalry", "unused to riding" or "without cavalry."

are also words that can not be described as other headwords, since some parts exist only in compounded or derived words.

This leaves a great portion of the morphological data from dictionaries insufficient for the sole basis of the morphological analysis in a database.

The 1% Sample

To further investigate the usefulness of morphological analyses of all headwords, I took a random one percent sample of headwords and evaluated whether they contained prefixes (alpha privatives counted separately), or were compounds. I began analysis for whether each word had derivations or was itself derived, but did not finish as I found the prefix and compound analysis compelling by itself.

The table below summarizes the findings. The total number of words analyzed was 1165. The fields below are not mutually exclusive, e.g. words counted for “compound and prefix” are also counted in “1% prefix” and “1% compound.”

Table 1: The 1% Sample

	total	**% of total**
1% alpha privative	**30**	**2.58%**
1% prefix	**436**	**37.42%**
1% compound	**315**	**27.04%**
are all 3	**0**	**0%**

alpha privative and prefix	**10**	**0.86%**
alpha privative and compound	**2**	**0.17%**
compound and prefix	**15**	**1.29%**
are 2 or more	**27**	**2.32%**

The number of prefixes and compounds suggests that around 60% or more of headwords in the LSJ can be described as combinations of other elements. There are some possible limitations to this count. Some words have parts that aren't headwords, others have parts that could be headwords but are not in the LSJ, and the actual frequency of the words themselves in texts is not accounted for here. Regardless, a figure of around 60%–not even accounting for derivational suffixes!–shows promise that the work is fruitful for producing useful data for the study of Greek word formation.

A Discussion of Working Definitions

Before going further, I set forth some working definitions of the kinds of morphemes present in Ancient Greek. While my specific purpose here is not to engage in such questions as "what is a compound" or "what is a prefix" in general or specific to Ancient Greek, a working answer to the question informs the structure of the database of Ancient Greek morphology and the numbers in my analysis above.

I look to the Cambridge Grammar of Classical Greek for some base working definitions:

> “Derivation: the addition of suffixes to a root to derive a new nominal
>
> or verbal form.”[13]
>
> “Composition: the combination of two (or more) nominal or verbal roots to form a new nominal or verbal form.”[14]
>
> “Prefixation: [forming a compound verb by] prefixing one or more prepositions (preverbs) to a simplex verb or a denominative verb.”[15]

By these definitions, when prepositions are compounded with nominals, they are not considered preverbs/prefixes.

However, it is often difficult to tell when an element was formed by prefixation or compounding. To take an example from the Cambridge Grammar of Classical Greek: *ἐπιστατ-έω* “preside”, which is formed from *ἐπι-στάτης* “president”, is sometimes treated as if it were formed by prefixation of *ἐπι-στατέω* (as evidenced by the use of the augment in finite past forms).[16] Treating each word so thoroughly would be time consuming and is outside of the current scope, even if it would be good to do eventually. Thus, pieces that can be preverbs in some words are generally considered preverbs in all words, except for compound adverbs such as *ἐπιπρό*.

More generally the question of “at what point in the formation of a word was each element added?” is difficult to answer for each word in a 116,460 word dictionary. While

[13] van Emde Boas et al. (2019), Section 23.1, p.260.
[14] van Emde Boas et al. (2019), section 23.1, p.260.
[15] van Emde Boas et al. (2019), section 23.51 p.276.
[16] van Emde Boas et al. (2019), section 23.51 p.276.

the LSJ does often give implicit morphological information here,[17] it does not always. Thus, I elect to give a more consistent agnostic approach to analysis, rather than make judgments on word formation order for words. This means splitting a word into its constituent parts unordered, which sometimes results in analyses that seem a little ridiculous: *ἄλαστος* “unforgetting” is analyzed as having an alpha privative, but so are all the words derived from it, e.g. *ἀλαστέω* “to be full of wrath.”

A list of what are counted as prefixes in this project: the preverbs – ἀνα-, ἀντι-, ἀπο-,δια- (ζα-), εἰσ- (ἐσ-), ἐκ-/ἐξ-, ἐν-,ἐπι-,κατα-, μετα- (ποδ-), παρα-,προ-, προσ-, συν- (ξυν-); other adverbial compounds εὐ-, δυσ-, παν-**,**. There are others counted as prefixes by the LSJ, including ἀρι-, ἀρχι-, να- (νη-), that I count as prefixes too, although I have not yet exhaustively annotated the data for less common ones.

The analysis of some words as either containing a prefix or being a compound is not always clear. Some prefixes, such as αὐτ- or ἀρχι- or παν-, appear to act as compounds in some cases and prefixes in others, as in the word *αὐτεπιστατέω***,** preceding a prefix with no connecting vowel where it is expected. I count *παντ-* in words as compounds, since it retains the tau of the stem, as in *παντ-άγαθος*, but it appears more commonly as *παν-*, where I count it as a prefix, as in *παν-άγαθος*. I have typically erred on classifying things as prefixes instead of compounds, although I am not entirely certain of these classifications.[18] Decisions of this sort should be straightforward to reverse if need be.

[17] It correctly splits *ἐπιστατ-έω*, implying that *-έω* is the most recent morpheme, rather than splitting it *ἐπι-στατέω.*

[18] For a general discussion of the difficulty differentiating these linguistically, see “Delineating Derivation and Compounding”, Olsen (2017), in Lieber, R., & Štekauer (2017).

Generally, any root combined with a connecting element (mostly omicron -*o*-) is considered a compound, and any root that is compounded without a connecting element is considered a prefix. Compounds with more than two roots are "multi-compounds." Compounds made up of multiple words or a phrase, (rather than the normal compounds composed of roots and a suffix), are considered multi-word or phrasal compounds.

Preliminary Work That Didn't Work

The (Insufficient) Suffix Algorithm and Trie Structure

The first idea to automate this process was to take a list of suffixes from Smyth's Greek Grammar and use it to analyze the suffixes present in each headword. This would create a base-suffix pair. To further check for accuracy of this base-suffix pair, new potential words would be composed by using the base and the other suffixes on the list, and then those words would be checked for in the LSJ.[19]

There are some limitations to this approach. One that may be obvious is that not all words have other words that are related by derivation in the LSJ, but might otherwise have obvious relations. This also includes a lack of ability to link internal stem changes, though that may be largely unavoidable in any method.

Another is that the method requires all suffixes, as well as concatenations of those suffixes, be manually defined as well as any morphophonological changes. The

[19] A description of the process is relegated to Appendix 1.

version that was tested did this by defining strings that might contain more than the suffix, such as the suffix *σις* being contained in the string *ξις*, but it was by no means exhaustive.

Results

Here are the results of a random 0.1% sample (n=116) of the outcome of the algorithm for stem-suffix pairs for a headword. Each was counted as “matched” (and so leading to a generated stem-suffix split) or “unmatched” (leading to no split), and then being counted as “true” if this was the correct split or no-split, or “false” if it was not. Theme vowels remaining on the stem were counted as incorrect.

Table 2: Suffix Algorithm Evaluation

Accuracy	**(true positives + true negatives)/all**	**77/116**	**0.664**
Precision	**true positives / (true positives + false positives)**	**35/52**	**0.673**
Recall	**true positives / (true positives + false negatives)**	**35/57**	**0.614**
F1 Score	**2(precision*recall) /(precision+recall)**		**0.042**

Also, in this analysis, I counted whether each headword had a compound or prefix (not distinguished in this count). 68 out of 116, or 58.6%, which lines up with the earlier 1% sample.

From the results, and flaws besides, I drew the conclusion that, either this needed a great deal of further refining to be any better than manual annotation, or to try a different approach.

BERT (Possible Uses)

A search for a better, less Ancient Greek specific method led me to BERT,[20] a natural language processing technique, and its WordPiece parser. BERT works by first using WordPiece to split words into a user-defined number of common sequences of characters, and then using those pieces to create mathematical representations of them. It does not require any manually annotated data for this, but annotated data is used to further refine the BERT model after this step.

I used BERT's WordPiece parser to tokenize the Greek headwords in the LSJ (normalized, both stripped of accents and not) to see whether it might produce splits that were morphologically meaningful. However, WordPiece isn't meant for fine-tuning to produce linguistically meaningful pieces of words; it is designed to prepare data for sentence level analysis, and the word pieces aren't required to be morphologically meaningful to be used for word piece embeddings and sentence analysis. It can produce accurate results, but it is not precise: when given *μονόχειρ*, the tokenizer produced *μονο-χειρ*, a linguistically meaningful split; but when given *χειρ*, it returned *χει-ρ*.

The results were mixed overall. In a 116 headword sample (~0.01% of the total), only 30 had derivational morphemes correctly split, but 71 had the correct split for

[20] Bidirectional Encoder Representations from Transformers (BERT). Devlin et al. (2019).

compounds or prefixes. Only 18 had both correct. This suggests that the WordPiece parser indeed cannot be used to consistently produce reliable, linguistically relevant results. Nevertheless, the fact that an unsupervised model might produce a split like *μονό-χειρ* suggests that a fine tuned supervised BERT model might be useful for more linguistically meaningful results.

With supervised training, BERT can be applied to do part-of-speech annotation for words in sentences (token classification), as well as sentence classification (text classification).[21] Hypothetically, this same process can be applied to characters in words–given some proper annotation of characters. Possible annotations might be "begin affix","end affix", "theme vowel", "begin compound piece","end compound piece". Similarly, overall classification might be possible, such as identifying words as compound or simplex.

The main reason I did not build any such model myself is the limits of my own technical understanding of how to build it. BERT models largely are intended to be used "off the shelf" and only fine tuned using a small amount of data, rather than spending time and massive amounts of processing power creating new models. There is a model by Brennan Nichcolson–Ancient Greek Char Bert[22]–which was built to predict characters in lacunae, rather than identify parts of a word. Perhaps it could be also used for morphological parsing somehow.

However, BERT may be of limited use for a dead language in which the majority of words have been identified and classified. Much of the annotation can be done

[21] Hugging Face (n.d.). https://huggingface.co/tasks/text-classification, https://huggingface.co/tasks/token-classification
[22] Nicholson (2020). https://github.com/brennannicholson/ancient-greek-char-bert

manually for a large base of vocabulary, and then expanded as necessary. Then, possibly, a BERT model could then be built using this annotated base data to help annotation of new words.

The Spreadsheet (Current Form of Data, and Methods)

After investigating these two options, I decided to opt for the simplest, if most arduous, option: hand annotation. The majority of this was done in tabular format in Microsoft Excel, derived from the XML data from Logeion. The headword splits from both the LSJ and Bailly 2020 were also used to aid the process. 68021 of 116460, or roughly 58.4%, of LSJ entries provide morphological analysis, and 35746 Bailly entries were matched to these headwords (~30.7% of LSJ entries).

The columns are first taken from what the Logeion XML explicitly labels: id, key,type, head, orthography[23] (split into part1 and part2). Gender and part of speech are implicit. In an attempt to mitigate this, I queried Morpheus with each headword in the LSJ in an attempt to annotate each for explicit labels for part of speech, gender, declension, stem type and derivation type, though not all headwords yielded results.

Beyond this, the annotation started in an *ad hoc* manner, so the schema has always been a work in progress. At first, alpha privatives and other prefixes were not separated, so some alpha privatives still remain in the “prefix” column. Some strings in the “theme vowel” are not in fact theme vowels. Bailly 2020 morphological data was also added to aid manual annotation partway through.

[23] This is how the morphological data of the LSJ is referred to in the Logeion data.

As it stands now: there are columns for annotation for whether it is an alpha privative, whether it has a prefix and whether it is a compound. This is done with a count of the number of characters associated with each, and the word string is separated into columns for these parts by excel formulae. For example, *ἀφηγηματικός (ἀφ-ηγ-η-ματ-ικός)* is annotated for as 0 in the alpha privative column, 2 in the prefix column for *ἀφ*, and 0 in the compound column. The formulae in other columns then calculate what string corresponds to the number. Some words have alpha privatives after other prefixes, and some have prefixes in the 2nd part of a compound. There are separate columns for these as well.

Accompanying these fields is a descriptor of the annotation “compound status” and “prefix status”, mostly marked as either “man” for “manual” and “auto” for “automatic”. This is somewhat misleading–it is all done by hand, but “man” typically indicates that the entry has been more thoroughly investigated, while “auto” indicates that it has had less thorough investigation if any. They are both susceptible to the error prone nature of hand annotation. Other annotations are hopefully self explanatory.

A majority of compounds are bipartite compounds, but tripartite or more compounds do exist. This is currently handled by splitting of the last element of these compounds and grouping the rest together. A boolean field ”multicompound” is marked TRUE if the compound is more than two parts. The limits of tabular data make this difficult to manage in a more granular way, since a new field would need to be available for each part.[24] Other words that are compounds formed from a concatenation of words

[24] The longest word in the LSJ is a massive compound from the end of Aristophanes’ *Ecclesiazusae*, and contains over 25 roots:
λοπαδοτεμαχοσελαχογαλεοκρανιολειψανοδριμυποτριμματοσιλφιοκαραβομελιτοκατακεχυμενοκιχλεπικοσσυφοφαττοπεριστεραλεκτρυονοπτοκεφαλλιοκιγκλοπελειολαγῳοσιραιοβαφητραγανοπτερύγων

such as *διότι* or *Διόσκοροι*, are marked with “multiword” in the ‘compound_status’ column. When connective elements contract with the 2nd part of the compound, the contracted vowel is typically kept in the 2nd part.

“Preverbs” that appear as primary components in compounds are treated as prefixes. To use a previous example, the *ἄφ-* in *ἄφιππος*, “unsuited for cavalry”, is treated as a prefix, even though the word is a compound of *ἀπό* and *ἱππ-*, rather than *ἵππος* being prefixed with *ἀπό*. This is just for procedural expediency, since otherwise it would need to be checked whether a word was derived from a verb that had one as a prefix, or a fresh compound word.

Alpha privatives is in a separate field from other prefixes. The separable “preverbs” are not in separate fields from inseparable prefixes. If an alpha privative has a clear corollary, it has been marked as an alpha privative, e.g. *φίλος* and *ἄφιλος*. Other times, the English definition has been relied on. However, there is not a one-to-one relationship between English definitions with “in-” or “un-” (or other indications of a privative) and Greek headwords with alpha privatives. Some headwords that have an alpha privative have definitions that obscure it. When the definition implies an alpha privative, however, I have typically accepted it.

Less populated fields include “inflected”, “dialect” “connected lemma”. If a word is inflected, “inflected” is set to TRUE and the headword–if one exists–that it is an inflection of is given in “connected lemma.” Similarly, dialect is given in the “dialect” column and the main LSJ headword given in “connected lemma”.

Less consistently, headwords are also annotated for suffixes. This includes theme vowel, suffixes (not including the most recent), and most recent suffix. Take again the word *ἀφηγηματικός (ἀφ-ηγ-η-ματ-ικός)*. It is labeled with 1 for the theme vowel *η*, 3 for suffixes *ματ*, and 4 for the most recent suffix *ικός*. This is the part of the annotations in the data that are least fleshed out.

Preliminary Results

Table 3: Preliminary results

	total	**% of total**	**% of annotated**	**1% sample (% of total)**
total a-priv (including post prefix alphas)	**4274**	**3.67**	**4.05**	**2.58**
total prefixes	**42081**	**36.13**	**36.81**	**37.42**
total compounds	**15620**	**13.41**	**26.09**	**27.04**
no prefix, no compound	**14693**	**12.62**	**-**	**36.14**
not fully annotated	**56924**	**48.88**	**-**	**-**
just a-priv	**2875**	**2.47**	**-**	**1.55**
just prefix	**40051**	**34.39**	**-**	**35.28**
just compound	**14171**	**12.17**	**-**	**25.58**
a-priv and prefix	**990**	**0.85**	**-**	**0.86**
a-priv and compound	**409**	**0.35**	**-**	**0.17**
compound and prefix	**1040**	**0.89**	**-**	**1.29**
Have 2 or More	**2439**	**2.09**	**-**	**2.32**
Have all 3	**3**	**0**	**-**	**0**

Not all headwords have been annotated, so the percentages don't add up to 100% (or more than 100%). Currently, over 98% are annotated for prefixes, 90% for alpha privatives, 51% for compounds, and less than 20% are annotated for suffixes.

Compared to the 1% sample, the prefixes are about the same: 36.1% to the sample's 37.4%. Similarly, we might expect the number of compounds to about double, as they make up 13.4% of a little over half the headwords, and the sample had 27%. "No prefix, no compound" is not explicitly annotated, but derived from words that are annotated explicitly indicating they have no prefix (not including alpha privatives), and are not compounds.

Future work

There remains much work to be done on the first pass of annotation. The annotations for prefixes, alpha privatives, compounds and suffixes need to be completed fully. Then the roots apparent from this need to be effectively grouped. Many improvements can be made for better matching and overall accuracy.

The underlying phonological representation of words is the end goal of representing headwords. Currently only the written form of words is represented, which do not account for derivational or inflectional morphophonological changes. Letters that represent two or more sounds can be undoubled, and morphophonological changes undone. Roots that change vowel grade with part of speech are not currently easily grouped. Parts of speech are still not annotated for all headwords, which is a necessary step towards deriving word formation rules from the data.

Then, there remains a lot of work to be done to get the data into a queryable database format. The current, unseparated state of prefixes and multiroot compounds need to be separated. A NoSQL database such as mongoDB, rather than a relational database such as SQL (MySQL, SQLite), seems preferable. The main disadvantage of a SQL database is the number of fields that would be required for this type of analysis. JSON format data, or a NoSQL database like mongodb would be able to handle the worst case scenarios. It also avoids a homonym problem that the Word Formation Latin project notes: with a SQL database, querying homonymic words can mix up which derivation comes from which base word. They offer solutions to the problem, but a non SQL format with that link explicit in each entry can perhaps skirt this entirely, at the cost of redundant data.

There are many further possible annotations as well. Semantic change that is not represented solely by a word's constituent morphemes is not currently represented in the data. Meanings can be endocentric (a meaning equal to the sum of its parts, or has a head constituent), or exocentric (holding a meaning different than the parts immediately suggest, lacks a head constituent)[25]. Moreover, the same word may have both endocentric and exocentric meanings. For instance, *διαβάλλω* means "throw or carry over or across" as *δια* and *βάλλω* might suggest, but those parts do not immediately suggest the additional meaning of "attack a man's character" or "slander". And many such words may skew towards one usage or another. Fields for "Has Endocentric Meaning" or "Has Exocentric Meaning" would make this more explicit. It would be useful to distinguish this systematically: words that are new to the reader can

[25] See Scalise & Bisetto "The Classification of Compounds" in Lieber, R., & Štekauer (2011).

be marked to show that the reader should or should not be able to extrapolate meaning based on the parts of an unknown word.

The order of composition might be made more explicit. This would bring it more in line with the way WFL is structured. Doing so requires more in depth annotation based on dictionaries and their implied composition order, as well as etymological dictionaries.

There are further semantic annotations that might be interesting as well: semantic annotation of morpheme usage, such as disambiguating which sense of a word is being used in a compound, or homonymic suffixes[26]; semantic classifications between the constituents of compounds; semantic categories such as plants, animals or medical uses of words.

Once it approaches a more mature state, the data could be further used to analyze a corpus. Thus far, I have only looked at the breadth of the lexicon, but not its frequency. What are the endocentric compounds that are most commonly used? Which derivational morphemes? Are the differences in derivational suffix usage significant across a work, a genre, or other corpus?

[26] e.g. -της in ναύτης “boater” (person concerned), κριτής “judger" (agent), from Smyth (1920), sections 839, 843, pages 229, 232

Conclusion

While the work is still incomplete, I believe the work so far shows why this database is desirable and how it might be useful pedagogically and academically. The 1% sample discussed shows that immensity of the lexicon can be reduced by as much as 60% by describing words in terms of their constituent parts, not even including suffixes. The incomplete data shows that this sample bears out. The annotations that I have started still need completion, there are theoretical obstacles to clear before the data is finalized, and beyond that there are yet more possible ways to annotate the data that could be useful, both for NLP techniques such as BERT and analysis of corpora.

Bibliography

Bailly, Anatole; Gréco, Gérard; Charbonnet, André; Wilde, Mark de; Maréchal, Bernard. *Dictionnaire grec-français*. (4th ed.) [online], 2020. http://gerardgreco.free.fr/spip.php?article52, https://latin-dict.github.io/dictionaries/Bailly2020.html, https://github.com/latin-dict/Bailly2020/releases.

Beekes, R. (2010). *Etymological dictionary of greek*. (A. Lubotsky, Ed.) (Vol. 1+2). Brill.

Chantraine, P. (1933). *La formation des noms en Grec Ancien*. Librairie Ancienne Honoré Champion.

Crane, Gregory. "Generating and Parsing Classical Greek." Literary and Linguistic Computing, vol. 6, no. 4, Jan. 1991, pp. 243–45, https://doi.org/10.1093/llc/6.4.243. https://github.com/PerseusDL/morpheus/.

Devlin, J; Chang, M; Lee, K, and Toutanova, K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In North American Association for Computational Linguistics (NAACL). https://arxiv.org/abs/1810.04805.

Hugging Face. (n.d.). *Tasks*. Hugging Face. Retrieved December 29, 2022, from https://huggingface.co/tasks.

Lieber, R., & Štekauer, P. (Eds.). (2017). *The Oxford Handbook of Derivational Morphology*. Oxford University Press.

> Olsen, S. (2017). Delineating Derivation and Compounding. In Lieber, R., & Štekauer, P. (Eds.) *The Oxford Handbook of Derivational Morphology* (pp. 26-49). Oxford University Press.

Lieber, R., & Štekauer, P. (Eds.). (2011). *The Oxford Handbook of Compounding*. Oxford University Press.

> Scalise, S., & Bisetto, A. (2011). The Classification of Compounds. In Lieber, R. & Štekauer, P. (Eds.), *The Oxford Handbook of Compounding* (pp. 34–53). Oxford University Press.

Liddel, H. G., Scott, R., & Jones, H. S. (1882). *A Greek-English Lexicon* (8th ed.). American Book Company. https://archive.org/details/greekenglishlexi00lidd_9, https://github.com/helmadik/LSJLogeion.

Litta, Eleonora. Passarotti, Marco. Culy, Chris (2018) *Word Formation Latin (WFL)*. https://doi.org/10.5281/zenodo.1492326. https://github.com/CIRCSE/WFL, https://wfl.marginalia.it/.

Nicholson, B. (2020). Ancient Greek Char Bert. GitHub. Retrieved 2022, from https://github.com/brennannicholson/ancient-greek-char-bert.

Smyth, H. W. (1920). *A Greek Grammar for Colleges*. American Book Company. https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0007.

Tribulato, O. (2015). *Ancient Greek Verb-Initial Compounds: Their diachronic development within the Greek compound system*. W. de Gruyter.

van Emde Boas, E., Rijksbaron, A., Huitink, L., & de Bakker, M. (2019). *The Cambridge Grammar of Classical Greek*. Cambridge University Press.

Appendix 1 - Suffix Algorithm in More Detail

Based on a user-defined (currently me) list of suffix strings:

-use a reverse-alphabetical ordered trie to determine what suffix strings are in a headword string. There may be multiple because of overlap, e.g. -ov, -ιον, -αδιον

-Iterating through that list of possible suffix strings,

-strip off the suffix string from word string: separate into a “root” and a “suffix”

-use the “root” to retrieve ***possible matches*** by using the alphabetical order trie

-from these possible matches, determine which possible suffixes *they* contain

-iterate through this new list of suffix strings

-add the suffix string to the original “root”

-look through ***possible matches*** for a match

-if there is a match, then a copy of that “root” is added to a list of possible “roots”

-the most common possible “root” is chosen

Appendix 2 - All the Data

The data, which constitutes a large portion of the work on this paper and project, is too much to print here, and is currently held in a GitHub repository:

https://github.com/barnes17/GreekWordFormation