A Nearest-Neighbor Approach to the Automatic Analysis of Ancient Greek Morphology

John Lee
Spoken Language Systems
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA
jsylee@csail.mit.edu

Abstract

We propose a data-driven method for automatically analyzing the morphology of ancient Greek. This method improves on existing ancient Greek analyzers in two ways. First, through the use of a nearest-neighbor machine learning framework, the analyzer requires no hand-crafted rules. Second, it is able to predict novel roots, and to rerank its predictions by exploiting a large, unlabelled corpus of ancient Greek.

1 Introduction

The civilization of ancient Greece, from which the Western world has received much of its heritage, has justly received a significant amount of scholarly attention. To gain a deeper understanding of the civilization, access to the essays, poems, and other Greek documents in the original language is indispensable.

Ancient Greek is a highly inflected Indo-European language[1]. A verb, for example, is inflected according to its person, number, voice, tense/aspect and mood. According to (Crane, 1991), “a single verb could have roughly 1,000 forms, and, if we consider that any verb may be preceded by up to three distinct prefixes, the number of forms explodes to roughly 5,000,000.” The inflections are realized by prefixes and suffixes to the stem, and sometimes spelling changes within the stem. These numerous forms can be further complicated by accents, and by additional spelling changes at morpheme boundaries for phonological reasons. The overall effect can yield an inflected form in which the root[2] is barely recognizable.

Indeed, a staple exercise for students of ancient Greek is to identify the root form of an inflected verb. This skill is essential; without knowing the root form, one cannot understand the meaning of the word, or even look it up in a dictionary.

For Classics scholars, these myriad forms also pose formidable challenges. In order to search for occurrences of a word in a corpus, all of its forms must be enumerated, since words do not frequently appear in their root forms. This procedure becomes extremely labor-intensive for small words that overlap with other common words (Crane, 1991).

Automatic morphological analysis of ancient Greek would be useful for both educational and research purposes. In fact, one of the first analyzers was developed as a pedagogical tool (Packard, 1973). Today, a widely used analyzer is embedded in the Perseus Digital Library (Crane, 1996), an internet resource utilized by both students and researchers.

This paper presents an analyzer of ancient Greek that infers the root form of a word. It introduces two innovations. First, *it utilizes a nearest-neighbor framework* that requires no hand-crafted rules, and provides analogies to facilitate learning.

[1] All Greek words are transcribed into the Roman alphabet in this paper. The acute, grave and circumflex accents are represented by diacritics, as in *ó*, *ò* and *õ*, respectively. Smooth breathing marks are omitted; rough breathing marks are signalled by *h*. Underbars used in *e̲* and *o̲* represent eta and omega.

[2] The root is also called the “base” or “lexical look-up” form, since it is the form conventionally used in dictionary entries. For verbs in ancient Greek, the root form is the first person singular present active indicative form. (cf. for English, it is the infinitive.) For nouns, it is the nominative singular form. For adjectives, it is the nominative singular masculine form.

CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning, pages 127–134
Manchester, August 2008

Person/Num	Form	Person/Num	Form
1st/singular	*lúo̲*	1st/plural	*lúomen*
2nd/singular	*lúeis*	2nd/plural	*lúete*
3rd/singular	*lúei*	3rd/plural	*lúousi(n)*

Table 1: Paradigm table for the present active indicative verb. It uses as example the verb *lúo̲* (“to loosen”), showing its inflections according to person and number.

Second, and perhaps more significantly, *it exploits a large, unlabelled corpus to improve the prediction of novel roots.*

The rest of the paper is organized as follows. We first motivate these innovations (§2) and summarize previous research in morphological analysis (§3). We then describe the data (§4) and our adaptations to the nearest-neighbor framework (§5-6), followed by evaluation results (§7).

2 Innovations

2.1 Use of Analogy and Nearest Neighbor

Typically, a student of ancient Greek is expected to memorize a series of “paradigms”, such as the one shown in Table 1, which can fill several pages in a grammar book. Although the paradigm table shows the inflection of only one particular verb, *lúo̲* (“to loosen”), the student needs to apply the patterns to other verbs. In practice, rather than abstracting the patterns, many students simply memorize these “paradigmatic” verbs, to be used as analogies for identifying the root form of an unseen verb. Suppose the unseen verb is *phéreis* (“you carry”); the reasoning would then be, “I know that *lúeis* is the second person singular form of the root *lúo̲*; similarly, *phéreis* must be the second person singular form of *phéro̲*.”

The use of analogy can be especially useful when dealing with a large number of rules, for example with the so-called “contract verbs”. The stem of a contract verb ends in a vowel; when a vowel-initial suffix is attached to the stem, spelling changes occur. For instance, the stem *ple̲ro-* (“to fill”) combined with the suffix *-omen* becomes *ple̲r-oũ-men*, due to interaction between two omicrons at the boundary. While it is possible to derive these changes from first principles, or memorize the rules for all vowel permutations (e.g., “*o*” + “*o*” = “*oũ*”), it might be easier to recall the spelling changes seen in a familiar verb (e.g., *ple̲róo̲* → *ple̲roũmen*), and then use analogy to infer the root of an unseen verb.

The nearest-neighbor machine learning framework is utilized to provide these analogies. Given a word in an inflected form (e.g., *phéreis*), the algorithm searches for the root form (*phéro̲*) among its “neighbors”, by making substitutions to its prefix and suffix. Valid substitutions are to be harvested from pairs of inflected and root forms (e.g., ⟨*lúeis*, *lúo̲*⟩) in the training set; these pairs, then, can serve as analogies to reinforce learning.

Furthermore, these affix substitutions can be learned automatically, reducing the amount of engineering efforts. They also increase the transparency of the analyzer, showing explicitly how it derives the root.

2.2 Novel Roots

Ancient Greek, in its many dialects, has been used from the time of Homer to the Middle Ages, in texts of a wide range of genres. Even the most comprehensive dictionaries do not completely cover its extensive vocabulary. To the best of our knowledge, all existing analyzers for ancient Greek require a pre-defined database of stems; thus, they are likely to run into words with unknown or novel roots, which they are not designed to analyze.

Rather than expanding an existing database to increase coverage, we create a mechanism to handle all novel roots. Since words do not often appear in their root forms, inferring a novel root from a surface form is no easy task (Lindén, 2008). We propose the use of unlabelled data to guide the determination of a novel root.

3 Previous Work

After a brief discussion on morphological analysis in general, we will review existing analyzers for ancient Greek in particular.

3.1 Morphological Analysis

A fundamental task in morphological analysis is the segmentation of a word into morphemes, that is, the smallest meaningful units in the word. Unsupervised methods have been shown to perform well in this task. In the recent PASCAL challenge, the best results were achieved by (Keshava and Pitler, 2006). Their algorithm discovers affixes by considering words that appear as substrings of other words, and by estimating probabilities for morpheme boundaries. Another successful approach is the use of Minimum Description Length, which iteratively shortens the length of the morphological grammar (Goldsmith, 2001).

Spelling changes at morpheme boundaries (e.g., *deny* but *deni-al*) can be captured by orthographic rules such as "change *y*- to *i*- when the suffix is -*al*". Such rules are specified manually in the two-level model of morphology (Koskenniemi, 1983), but they can also be induced (Dasgupta, 2007). Allomorphs (e.g., "*deni*" and "*deny*") are also automatically identified in (Dasgupta, 2007), but the general problem of recognizing highly irregular forms is examined more extensively in (Yarowsky and Wicentowski, 2000). They attempt to align every verb to its root form, by exploiting a combination of frequency similarity, context similarity, edit distance and morphological transformation probabilities, all estimated from an unannotated corpus. An accuracy of 80.4% was achieved for highly irregular words in the test set.

3.2 Challenges for Ancient Greek

Ancient Greek presents a few difficulties that prevent a naive application of the minimally supervised approach in (Yarowsky and Wicentowski, 2000). First, frequency and context analyses are sensitive to data sparseness, which is more pronounced in heavily inflected languages, such as Greek, than in English. Many inflected forms do not appear more than a few times. Second, many root forms do not appear[3] in the corpus. In Finnish and Swahili, also highly inflected languages, only 40 to 50% of words appear in root forms (Lindén, 2008). The same may be expected of ancient Greek.

Indeed, for these languages, predicting novel roots is a challenging problem. This task has been tackled in (Adler et al., 2008) for modern Hebrew, and in (Lindén, 2008) for Finnish. In the former, features such as letter n-grams and word-formation patterns are used to predict the morphology of Hebrew words unknown to an existing analyzer. In the latter, a probabilistic approach is used for harvesting prefixes and suffixes in Finnish words, favoring the longer ones. However, no strategy was proposed for irregular spelling in stems.

Surface Form	Morphological Annotation	Root Form
kaì (*and*)	Conjunction	*kaí*
pneũma (spirit)	Noun 3rd decl	*pneũma*
theoũ (God)	Noun 2nd decl	*theós*
epephéreto (hover)	Verb	*phéro̲*

Table 2: Sample data from parts of Genesis 1:2 ("and the Spirit of God was hovering over ..."). The original annotation is more extensive, and only the portion utilized in this research is shown here.

3.3 Ancient Greek Morphological Analysis

The two most well-known analyzers for ancient Greek are both rule-based systems, requiring *a priori* knowledge of the possible stems and affixes, which are manually compiled. To give a rough idea, some 40,000 stems and 13,000 inflections are known by the Morpheus system, which will be described below.

The algorithm in Morph (Packard, 1973) searches for possible endings that would result in a stem in its database. If unsuccessful, it then attempts to remove prepositions and prefixes from the beginning of the word. Accents, essential for disambiguation in some cases, are ignored. The analyzer was applied on Plato's *Apology* to study the distribution of word endings, for the purpose of optimizing the order of grammar topics to be covered in an introductory course. Evaluation of the analyzer stressed this pedagogical perspective, and the accuracy of the analyses is not reported.

Morpheus (Crane, 1991) augments Morph with a generation component which, given a stem, enumerates all possible inflections in different dialects, including accents. When accents are considered during analysis, the precision of the analyzer improves by a quarter. However, the actual precision and the test set are not specified.

In this paper, we have opted for a data-driven approach, to automatically determine the stems and affixes from training data.

4 Data

4.1 Morphology Data

We used the Septuagint corpus[4] prepared by the Center for Computer Analysis of Texts at the University of Pennsylvania. The Septuagint, dating from the third to first centuries BCE, is a

[3] The root forms of contract verbs, e.g. *ple̲róo̲*, are not even inflected forms.

[4] http://ccat.sas.upenn.edu/gopher/text/religion/biblical/

Part-of-speech	Percent
Verbs	68.6%
Adjectives	10.4%
Nouns (1st declension)	5.6%
Nouns (2nd declension masculine)	4.3%
Nouns (2nd declension neuter)	2.8%
Nouns (3rd declension)	7.6%
other	0.7%

Table 3: Statistics on the parts-of-speech of the words in the test set, considering only unique words.

Greek translation of the Hebrew Bible. The corpus is morphologically analyzed, and Table 2 shows some sample data.

The corpus is split into training and test sets. The training set is made up of the whole Septuagint except the first five books. It consists of about 470K words, with 37,842 unique words. The first five books, also known as the Torah or Pentateuch, constitute the test set. It contains about 120K words, of which there are 3,437 unique words not seen in the training set, and 7,381 unique words seen in training set. A breakdown of the parts-of-speech of the test set is provided in Table 3. Proper nouns, many of which do not decline, are excluded from our evaluation.

4.2 Unlabelled Data

To guide the prediction of novel roots, we utilize the *Thesaurus Linguae Graecae* (Berkowitz and Squitter, 1986) corpus. The corpus contains more than one million unique words, drawn from a wide variety of ancient Greek texts.

4.3 Evaluation

Many common words in the test set are also seen in the training set. Rather than artificially boosting the accuracy rate, we will evaluate performance on unique words rather than all words individually.

Some surface forms have more than one possible root form. For example, the word *purõ̲n* may be inflected from the noun *purá* (“altar”), or *purós* (“wheat”), or *pũr* (“fire”). It would be necessary to examine the context to select the appropriate noun, but morphological disambiguation (Hakkani-Tür et al., 2002) is beyond the scope of this paper. In these cases, legitimate root forms proposed by our analyzer may be rejected, but we pay this price in return for an automatic evaluation procedure.

5 Nearest-Neighbor Approach

The memory-based machine learning framework performs well on a benchmark of language learning tasks (Daelemans, 1999), including morphological segmentation of Dutch (van den Bosch, 1999). In this framework, feature vectors are extracted from the training set and stored in a database of instances, called the *instance base*. A distance metric is then defined. For each test instance, its set of nearest neighbors is retrieved from the instance base, and the majority label of the set is returned.

We now adapt this framework to our task, first defining the distance metric (current section), then describing the search algorithm for nearest neighbors (§6).

5.1 Distance Metric

Every word consists of a stem, a (possibly empty) prefix and a (possibly empty) suffix. If two words share a common stem, one can be transformed to the other by substituting its prefix and suffix with their counterparts in the other word. We will call these substitutions the *prefix transformation* and the *suffix transformation*.

The “distance” between two words is to be defined in terms of these transformations. It would be desirable for words that are inflected from the same root to be near neighbors. A distance metric can achieve this effect by favoring prefix and suffix transformations that are frequently observed among words inflected from the same root. We thus provisionally define “distance” as the sum of the frequency counts of the prefix and suffix transformations required to turn one word to the other.

5.2 Stems and Affixes

Defining “Stem” To count the frequencies of prefix and suffix transformations, the stem of each word in the training set must be determined. Ideally, all words inflected from the same root should share the same stem. Unfortunately, for ancient Greek, it is difficult to insist upon such a common stem. In some cases, the stems are completely different[5]; in others, the common stem is obfuscated

[5] Each verb can have up to six different stems, known as the “principal parts”. In extreme cases, a stem may appear completely unrelated to the root on the surface. For example, *oíso̲* and *é̲negkon* are both stems of the root *phéro̲* (“to carry”). A comparable example in English is the inflected verb form *went* and its root form *go*.

Word	Prefix	Stem	Suffix		Prefix Transformation	Suffix Transformation
(root) *lúo̲*	-	*lú*	*o̲*	(root,1)	ϵ ↔ `e`	`o̲` ↔ `eto`
(1) *elúeto*	*e*	*lú*	*eto*	(root,2)	ϵ ↔ `para`	`o̲` ↔ `sai`
(2) *paralũsai*	*para*	*lũ*	*sai*	(root,3)	ϵ ↔ `ek`	`o̲` ↔ `thé̲sontai`
(3) *ekluthé̲sontai*	*ek*	*lu*	*thé̲sontai*	(1,2)	`e` ↔ `para`	`eto` ↔ `sai`
				(1,3)	`e` ↔ `ek`	`eto` ↔ `thé̲sontai`
				(2,3)	`para` ↔ `ek`	`sai` ↔ `thé̲sontai`

Table 4: The verb root *lúo̲* ("to loosen") and three of its inflected forms are shown. Each inflected form is compared with the root form, as well as the other inflected forms. The "stem", defined as the longest common substring, is determined for each pair. The prefix and suffix transformations are then extracted. ϵ represents the empty string.

in surface forms due to spelling changes[6].

We resort to a functional definition of "stem" — the longest common substring of a *pair* of words. Some examples are shown in Table 4.

Refinements to Definition Three more refinements to the definition of "stem" have been found to be helpful. First, accents are ignored when determining the longest common substring. Accents on stems often change in the process of inflection. These changes are illustrated in Table 4 by the stem *lu*, whose letter *u* has an acute accent, a circumflex accent, and no accent in the three inflected forms.

Second, a minimum length is required for the stem. On the one hand, some pairs, such as *ágo̲* ("to lead") and *áxo̲*, do have a stem of length one ("*a*"). On the other hand, allowing very short stems can hurt performance, since many spurious stems may be misconstrued, such as "*e*" between *phéro̲* and *é̲negkon*. The minimum stem length is empirically set at two for this paper.

Length alone cannot filter out all spurious stems. For example, for the pair *patéo̲* ("to walk") and an inflected form *katepáte̲san*, there are two equally long candidate stems, **ate* and *pat*. The latter yields affixes such as "*-éo̲*" and "*-e̲san*", which are relatively frequent[7]. On this basis, the latter stem is chosen.

Some further ways to reduce the noise are to require an affix transformation to occur at least a minimum number of times in the training set, and to restrict the phonological context in which the transformation can be applied[8]. While significantly reducing recall, these additional restrictions yield only a limited boost in precision.

6 Algorithm

In the training step, a set of prefix and suffix transformations, along with their counts, is compiled for each part-of-speech. These counts enable us to compute the distance between any two words, and hence determine the "nearest neighbor" of a word.

At testing, given an inflected form, its neighbor is any word to which it can be transformed using the affix transformations. We first try to find its nearest neighbor in the training set (§6.1); if no neighbor is found, a novel root is predicted (§6.2).

6.1 Finding Known Roots

If the input word itself appears in the training set, we simply look up its morphological analysis.

If the input word is not seen in the training set, its root form or another inflected form may still be found. We try to transform the input word to the nearest such word, i.e., by using the most frequent prefix and suffix transformations, according to the distance metric (§5.1).

Irregular Stem Spelling Typically, if there are no spelling changes in the stem, the input word can be transformed directly to the root, e.g., from *phéreis* to *phéro̲*. If the spelling of the stem is substantially different, it is likely to be transformed to another inflected form of the root that contains the same irregular stem. For example, the word *prosexé̲negken* bears little resemblance to its root *phéro̲*, but it can be mapped to the word *é̲negken*

[6] For example, the stem *oz* in the root form *ózo̲* ("to smell") is changed to *o̲s* in *exó̲sthe̲san*, an aorist passive form.

[7] The frequency of each affix is counted in a preliminary round, with each affix receiving a half count in cases of tied stem length.

[8] For example, a certain suffix transformation may be valid only when the stem ends in certain letters.

in the training set, from which we retrieve its root form *phéro̲*.

Search Order Some affixes are circumfixes; that is, both the prefix and the suffix must occur together. For example, the suffix *-eto* cannot be applied on its own, but must always be used in conjunction with the prefix *e-*, to form words such as *elúeto*, as shown in Table 4.

Other affixes, however, can freely mix with one another, and not all combinations are attested in the training set. This is particularly common when the prefix contains two or more prepositions. For example, the combination *dia-kata-* occurs only two times in the training set, but it can potentially pair with a large number of different suffixes.

Hence, the search for neighbors proceeds in two stages. In the first stage (denoted CIRCUMFIX), the search is restricted to circumfixes, that is, requiring that at least one word-pair in the training set contain both the prefix and suffix transformations. This restriction is prone to data sparseness; if no neighbor is found, the prefix and suffix transformations are then allowed to be applied separately in the second stage (denoted PREFIX/SUFFIX).

6.2 Proposing Novel Roots

A word may be derived from a root of which no inflected form is seen in the training set. Naturally, no neighbor would be found in the previous step, and a novel root must be proposed. We apply the prefix and suffix transformations learned in §5.2, using only circumfixes observed between an inflected form and a root form. For obvious reasons, the resulting string is no longer required to be a neighbor, i.e., a word seen in the training set.

Typically, the various transformations produce many candidate roots. For example, the word *homome̲tríou* ("born of the same mother"), a masculine genitive adjective, can be transformed to its root adjective *homomé̲trios*, but it could equally well be transformed into a hypothetical neuter noun, **homomé̲trion*. Both are perfectly plausible roots.

The automatically discovered affix transformations inevitably contain some noise. When dealing with known roots, much of the noise is suppressed because misapplications of these transformations seldom turn the input word into a real word found in the training set. When proposing novel roots, we no longer enjoy this constraint. Although the distance metric still helps discriminate against invalid candidates, the increased ambiguity leads to lower accuracy. We address this issue by exploiting a large, unlabelled corpus.

Use of Unlabelled Corpus If a proposed root form is correct, it should be able to generate some inflected forms attested in a large corpus. Intuitively, the "productivity" of the root form may correlate with its correctness.

To generate inflected forms from a root, we simply take the set of affix transformations observed from inflected forms to roots, and reverse the transformations. Continuing with the above example, we generate inflected forms for both candidate roots, the adjective *homomé̲trios*, and the hypothetical neuter noun **homomé̲trion*. While a few inflected forms are generated by both candidates, three are unique to the adjective — *homomé̲trios*, *homomé̲trioi* and *homomé̲trian* — the nominative masculine singular and plural, and the accusative feminine singular, respectively. None of these could have been inflected from a neuter noun.

A straightforward notion of "productivity" of a root would be simply the number of inflected forms attested in the large corpus. It can be further refined, however, by considering the prevalence of the inflected forms. That is, a form generated with more common affix transformations should be given greater weight than one generated with less common ones. Suppose two candidate roots, the adjective *telesphóros* ("bringing to an end") and the hypothetical verb **telesphoróo̲*, are being considered. Both can generate the inflected form *telesphórou*, the former as the masculine genitive adjective, and the latter as either an imperfect indicative or present imperative contract verb. Since the inflection of the adjective is more frequent in the training set than that of the relatively rare class of contract verbs, the existence of *telesphórou* should lend greater weight to the adjective.

Hence, the "productivity" metric of a novel root is the number of words in the large corpus that it can generate with affix transformations, weighted by the frequencies of those transformations.

7 Experiments

Some statistics on the test set are presented in Table 3. Of the 7,381 words that are seen in the training set, 98.2% received the correct root form. The

Transformation Type	Proportion	Accuracy
Circumfix	77.5%	94.5%
Prefix/Suffix	10.8%	61.2%
Novel Roots	11.7%	50.0%
Overall	100%	85.7%

Table 5: After excluding known words, which attain an accuracy of 98.2%, the performance on the remaining 3437 unique words in the test set is shown above. Please see §7 for discussions. Results for novel roots are presented in further detail in Table 6.

Evaluation Method	Accuracy
Baseline	45.0%
TLG Rerank	50.0%
+Ignore accents	55.2%
+Oracle POS	65.5%

Table 6: Results for predicting novel roots, for the 402 words for whom no neighbor was found. Baseline uses the distance metric (§5.1) as before; TLG Rerank exploits the unlabelled Thesaurus Linguae Graecae corpus to re-rank the top candidates (§6.2) proposed by Baseline.

remaining 1.8% had multiple possible roots; an examination of the context would be needed for disambiguation (see comments in §4.3).

Table 5 presents the accuracy of the predicted roots, after excluding the 7,381 seen words. The result is broken down according to the type of transformation; for the "Novel Roots" type, more detailed results are presented in Table 6.

As discussed in §6.1, the algorithm first searched with Circumfix. For 77.5% of the words, a neighbor was found using this subset of affix transformations. The rest were then processed using the back-up procedure, Prefix/Suffix, allowing prefix and suffix transformations culled from different word-pairs. This procedure found neighbors for 10.8% of the words; novel roots were hypothesized for the remainder.

Not surprisingly, known roots were more reliably predicted (94.5%) with circumfixes than with separate prefixes and suffixes (61.2%), but both categories still achieved higher accuracy than the challenging task of proposing novel roots (50.0%). We now take a closer look at the errors for both known and novel roots.

7.1 Known Roots

There are three main sources of error. The first is noise in the affix transformations. For example, the spurious prefix transformation p↔ph was derived from the pair *phéro̲* and *perienḗ̲gkasan*. When applied on *pasáto̲*, along with a suffix transformation, it yielded the false root form *phásko̲*.

A second source can be attributed to incorrect affix boundaries. For example, *ekteínantes* was misconstrued as having "*e-* " rather than the preposition *ek* as prefix. This prefix is by itself perfectly viable, but "*e-*" and "*-antes*" cannot occur together as a circumfix. The resulting string happened to match the root *kteíno̲*, rather than the true root *teíno̲*.

A third source is confusion between parts-of-speech, most commonly noun and verb. For example, the nearest neighbor of the genitive noun *lupõ̲n* was the verb *lupé̲sei*, yielding the verb root *lupéo̲* rather than the noun *lúpe̲*.

7.2 Novel Roots

As a baseline, the distance metric (§5.1) was used alone to rank the novel candidate roots. As seen in Table 6, performance dropped to 45.0%.

When the Thesaurus Linguae Graecae corpus was utilized to rerank the novel candidate roots proposed by the baseline, an absolute gain[9] of 5% was achieved. A further 5.2% of the mistakes were due to placing the accent incorrectly, such as *kte̲nótrophos* rather than *kte̲notróphos*, mostly on nouns and adjectives. These mistakes are difficult to rectify, since multiple positions are often possible[10].

Finally, to measure the extent to which part-of-speech (POS) confusions are responsible, we performed an experiment in which the gold-standard POS of each word was supplied to the analyzer (see "Oracle POS" in Table 6). When deriving novel roots, only those affix transformations belonging to the oracle POS were considered. With this constraint, accuracy rose to 65.5%.

[9] The significance level is at $p = 0.11$, according to McNemar's test. The improvement is not statistically significant, and may be a reflection of the relatively small test set.

[10] The accent in an inflected noun retains its position in the root, unless that position violates certain phonological rules. In many cases, there is no reliable way to predict the accent position in the root noun from the position in the inflected form.

8 Conclusion

We have proposed a nearest-neighbor machine learning framework for analyzing ancient Greek morphology. This framework is data-driven, with automatic discovery of stems and affixes. The analyzer is able to predict novel roots. A significant novelty is the exploitation of a large, unlabelled corpus to improve performance.

We plan to further improve the derivation of novel roots by predicting their parts-of-speech from context, and by incorporating distributional information (Yarowsky and Wicentowski, 2000).

Acknowledgments

The author would like to thank Stephanie Seneff, Kalliroi Georgila, Konstantinos Katsiapis and Steven Lulich for their insightful comments.

References

Meni Adler, Yoav Goldberg, David Gabay, and Michael Elhadad. 2008. Unsupervised Lexicon-based Resolution of Unknown Words for Full Morphological Analysis. *Proc. ACL*. Columbus, OH.

Luci Berkowitz and Karl A. Squitter. 1986. *Thesaurus Linguae Graecae*. Oxford University Press, UK.

Antal van den Bosch and Walter Daelemans. 1999. Memory-based Morphological Analysis. *Proc. ACL*. College Park, MD.

Gregory Crane. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Gregory Crane. 1996. *Perseus 2.0: Interactive Sources and Studies on Ancient Greece*. Yale University Press, New Haven, CT.

Walter Daelemans, Antal van den Bosch and Jakub Zavrel. 1999. Forgetting Exceptions is Harmful in Language Learning. *Machine Learning*, 34:11–41.

Sajib Dasgupta and Vincent Ng. 2007. High-Performance, Language-Independent Morphological Segmentation. *Proc. HLT-NAACL*. Rochester, NY.

John Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

Dilek Z. Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2002. Statistical Morphological Disambiguation for Agglutinative Languages. *Computers and the Humanities*, 36(4):381–410.

Samarth Keshava and Emily Pitler. 2006. A Simpler, Intuitive Approach to Morpheme Induction. *Proc. 2nd PASCAL Challenges Workshop*. Venice, Italy.

Kimmo Koskenniemi. 1983. Two-level morphology: a general computation model for word-form recognition and production. *Publication No. 11, Department of General Linguistics, University of Helsinki*. Helsinki, Finland.

Krister Lindén. 2008. A Probabilistic Model for Guessing Base Forms of New Words by Analogy. *Proc. CICLing*. Haifa, Israel.

David W. Packard. 1973. Computer-assisted Morphological Analysis of Ancient Greek. *Proc. 5th Conference on Computational Linguistics*. Pisa, Italy.

David Yarowsky and Richard Wicentowski. 2000. Minimally Supervised Morphological Analysis by Multimodal Alignment. *Proc. ACL*. Hong Kong, China.