Topic modelling experiments on Hellenistic corpora

Ryder Wishart
McMaster Divinity College
Hamilton, Ontario
Canada
ryderwishart@gmail.com

Prokopis Prokopidis
Institute for Language and Speech Processing/Athena RIC
Athens, Greece
prokopis@ilsp.gr

Abstract

The focus of this study is Hellenistic Greek, a variation of Greek that continues to be of particular interest within the humanities. The Hellenistic variant of Greek, we argue, requires tools that are specifically tuned to its orthographic and semantic idiosyncrasies. This paper aims to put available documents to use in two ways: 1) by describing the development of a POS tagger and a lemmatizer trained on annotated texts written in Hellenistic Greek, and 2) by representing the lemmatized products as topic models in order to examine the effects of a) automatically processing the texts, and b) semi-automatically correcting the output of the lemmatizer on tokens occurring frequently in Hellenistic Greek corpora. In addition to topic models, we also generate and compare lists of semantically related words.

1 Introduction and Motivation

Research into the ancient Greek language and culture has been a cornerstone of western humanities, specifically within the domain of classics. Research into classics typically examines writers such as Homer, Plato, and Aristotle, who have served as important influences in the formation and development of Western culture. However, the Greek language did not evolve directly from the classical period into the modern period; there are at least three important milestones in between: Hellenistic, Byzantine, and Medieval Greek. The focus of this study is Hellenistic Greek, as this variation of Greek, sometimes called "Koine Greek" has been of particular interest within the humanities. The reason for this interest is that Hellenistic Greek incorporates the texts of the New Testament and the Septuagint (or LXX, i.e. the Greek translation of the Hebrew Scriptures). The Hellenistic variant of Greek, we would argue, requires tools that are specifically tuned to its orthographic and even semantic idiosyncrasies.[1]

While there have been significant advances in the use of language technology (LT) for Classical Greek,[2] very little has been attempted in the form of LT for Hellenistic Greek texts or language, though more is currently being explored in this direction. One of the most prominent endeavours in this direction is the OpenText.org project, which provides a freely available annotated New Testament.[3] OpenText is unique in that it annotates discourse features such as clause and word groups. Current initiatives that

[1] For example, there are numerous spelling variations that can be observed when comparing Hellenistic and Classical Greek. As well, comparing these two variants of Greek, one can observe the expected diachronic shift not only in word usage, but also grammatical usage. Thus, for the study of Hellenistic Greek, tools tuned to its conventions are necessary.

[2] For example, see the Perseus project at http://www.perseus.tufts.edu/hopper/. Another project is the Classical Language Toolkit (CLTK), which builds off of the Natural Language Toolkit in order to develop tools for the analysis of ancient languages, including Classical Greek.

[3] Another initiative is the Open Greek and Latin Project (http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/), which aims to provide a free archive of every Greek and Latin source from the earliest extant up to the beginning of the seventh century. See also the Perseus database (https://github.com/PerseusDL).

provide digital resources for the study of Hellenistic Greek have chiefly aimed at providing new resources and corpora. This paper, by contrast, aims to put available documents to use in two ways: 1) by developing a POS tagger and a lemmatizer trained on annotated texts written in Hellenistic Greek, and 2) by representing the lemmatized products as topic models in order to examine the effects of a) automatically processing the texts, and b) correcting the output of the lemmatizer on tokens occurring frequently in Hellenistic Greek corpora. In addition to topic models, we also generate and compare lists of semantically related words. In order to substantiate our claim that tools tailored to the Hellenistic variety in particular are necessary, as opposed to simply Ancient Greek data that is currently available through Perseus, we train an Ancient Greek lemmatizer+POS tagger and provide a comparison of the number of errors that occur in a small subset of our test corpus.

The motivation for this study is twofold. First, we are convinced that automatic or semi-automatic means of processing linguistic data will become increasingly important in the digital humanities. Secondly, we recognize that, without preprocessing of texts, corpus linguistics is significantly handicapped in its ability to furnish data and insights into highly inflected and non-configurational languages such as Greek.

2 Description of Dataset

There are two types of data that we used in our study: two annotated training corpora, and a non-annotated test corpus that resembles the representative corpus of Hellenistic Greek suggested by O'Donnell [5]. The first training corpus we used was an annotated version of the Greek New Testament, the SBL (Society of Biblical Literature) Greek New Testament.[4] In the original annotated version, punctuation marks were not separated from preceding tokens. We used simple heuristics to normalize such tokens and converted the resource into the CoNLL format, with empty lines separating sentences. The final training corpus contained approximately 140K tokens in 10.5K sentences. After using this annotated corpus to train our POS tagger and lemmatizer, we then used the generated models to annotate the test corpus. Because the encodings of the training and the test corpora differed,[5] we converted the accented vowels in the test corpus from Extended Greek to Greek and Coptic. The second training corpus, used for performance comparison with the SBLGNT (see section 6.4 below) was the Ancient Greek Dependency Treebank (AGDT) v. 1.7.[6] This corpus contained approximately 355K tokens in 24.8K sentences.

The test corpus, while based on O'Donnell's representative corpus, is different in two non-trivial aspects. On the one hand, two significant domains of Hellenistic Greek, the documentary papyri as well as the inscriptions, are missing from our corpus. On the other hand, O'Donnell had capped the length of certain documents such as the works of Strabo, Polybius, and Arrian, at either 20000 or 30000 words.[7] We did not cap the length of these documents, however, as we were concerned with accuracy of the lexical data, not necessarily a balanced representation of genera. Thus, the test corpus ends up being significantly larger than O'Donnell's originally suggested corpus, but the resulting token/lemma data—the data that is used to train and manually correct the lemmatizer—is not distorted in the same way that the semantic content of the topic models is.[8] Put differently, the topic models cannot accurately describe the semantics of the language variety without a balanced corpus, because some text types will be over- or under-represented. By contrast, the token/lemma data relies only on individual tokens, without respect to the larger discourse units of the corpus. Our test corpus contained approximately 1.81M tokens in 91K sentences. We also used a subset of this corpus, comprised of 100 sentences taken from three works in the target variety of Greek in order to test the performance of the SBLGNT versus the AGDT (see section 6.4 below). See Table 1 for a side-by-side comparison of the corpora used in this study.

[4] This annotation, the MorphGNT, is actually a compilation of several annotated Greek New Testaments (https://github.com/morphgnt/sblgnt)

[5] Because ancient and modern Greek differ in the accents they use, the encodings of the accents are often problematic. For more discussion about this issue, visit https://wiki.digitalclassicist.org/Greek_Unicode_duplicated_vowels

[6] The AGDT (https://perseusdl.github.io/treebank_data/) is currently in v. 2.1, but 1.7 was used for this test in order to keep the total size of the training corpora relatively closer.

[7] This length, though somewhat arbitrary (see discussion in O'Donnell [5]), is intended to keep particular subvarieties of Hellenistic Greek such as literary or Atticistic from dominating the data. For discussion of O'Donnell's corpus, see Pang [6].

[8] See discussion on the topic models below.

SBLGNT	AGDT	O'Donnell (non-truncated)
New Testament (CE 1) Matthew: 18556 words Mark: 11424 words Luke: 19696 words John: 15763 words Acts: 18687 words Romans: 7199 words 1 Corinthians: 6895 words 2 Corinthians: 4542 words Galatians: 2255 words Ephesians: 2457 words Philippians: 1645 words Colossians: 1597 words 1 Thessalonians: 1500 words 2 Thessalonians: 831 words 1 Timothy: 1617 words 2 Timothy: 1264 words Titus: 682 words Philemon: 342 words Hebrews: 5054 words James: 1765 words 1 Peter: 1709 words 2 Peter: 1121 words 1 John: 2160 words 2 John: 249 words 3 John: 222 words Jude: 465 words Revelation: 9918 words	**Hesiod (BCE 8?)** Shield of Heracles: 3834 words Hesiod, Theogony: 8106 words Hesiod, Works and Days: 6941 words **Homer (BCE 8)** Iliad: 128102 words Odyssey: 104467 words **Aeschylus (BCE 6-5)** Agamemnon: 9806 words Eumenides: 6380 words Libation Bearers: 6566 words Persians: 6270 words Prometheus Bound: 7058 words Seven Against Thebes: 6232 words Suppliants: 5949 words **Sophocles (BCE 5)** Ajax: 9474 words Antigone: 8751 words Electra: 10489 words Oedipus Tyrannus: 11185 words Trachiniae: 8822 words **Plato (BCE 5-4)** Euthyphro: 6097 words	**Plutarch (CE 1-2)** Cato Minor: 17031 words **Philo (BCE 1—CE 1)** On the Creation: 31852 words **New Testament** (see SBLGNT) **Diodorus Siculus (CE 1)** Bibliotheca Historica: 417681 words **Strabo (BCE 1—CE 1)** Geographica: 298655 words **Cassius Dio (CE 2-3)** Historiae Romanae: 379170 words **Josephus (CE 1)** Life: 16224 words **LXX (BCE 3—CE 3)** Judges: 16324 words 2 Esdras: 13618 words Tobit: 7421 words **Polybius (BCE 3-2)** Historiae: 326081 words **Pseudo-Apollodorus (CE 1-2)** Bibliotheca: 28249 words **Epictetus (CE 1-2)** Dissertationes: 78165 words **Didache (CE 2)** 2241 words **Shepherd of Hermas (CE 2)** 27819 words **Ignatius (CE 1-2)** Ephesians: 7956 words
Total: 139615 words[9]	**Total:** 354529 words	**Total:** 1808102 words

Table 1. Corpora, centuries, and word counts.

In summary, then, the annotated Greek New Testament provided us with enough data of the targeted variety to train the lemmatizer up to a point where it became feasible to provide manual corrections,[10] as illustrated by the improvement of the lexical data that resulted from manually correcting high-frequency token/lemma pairs.

3 Pre- and post-processing tools, training, and application

As described in the previous section, we first used the SBL Greek New Testament to train the POS tagger and lemmatizer, as the Greek New Testament falls directly in the middle of the era of Hellenistic Greek. For training the POS tagger and the lemmatizer, we used the MarMoT+Lemming toolkit [4]. We then attempted to POS tag and lemmatize the O'Donnell corpus, with reasonably good results.

[9] Word counts of the NT will vary more than other ancient texts due to the proliferation of critical editions and the large number of extant manuscripts. Our word count here includes book headings and several other miscellaneous insertions, thus skewing the total by several hundred tokens. These tokens are ideally lemmatized or else filtered out after post-processing as having the POS tag "NONE".

[10] We manually corrected the tokens that appeared >1000 times, which comprised the first 1574 rows of the token/lemma table. At this point, it was not feasible to manually correct more data, as the total number of rows was just below 1M, and the decreasing frequency of the tokens on each row would have had a quickly diminishing return for the time spent.

In order to track our progress, we used this first attempt to generate some topic models using the open source software GenSim [7], and we used pyLDAvis[11] to visually represent the topic models.

Because pyLDAvis can be used to display high frequency lemmas in the data, using it allowed us to more quickly assess where errors appeared frequently in the lemmatized data. Even though there are presumably thousands of unique and erroneous token/lemma pairs in our data, only high-frequency errors were evident.

In order to generate the topic models, we produced a lemmatized version of the test corpus, where every word was replaced by its lemma. We then filtered out parts of speech corresponding to non-content tokens like punctuation marks, numbers, articles, pronouns, etc,. Next we used GenSim to train a Latent Dirichlet Allocation [1], or LDA, model of the corpus, and then represented the resulting corpus, dictionary, and model using pyLDAvis, which can be saved as an HTML file and viewed through an internet browser.

4 Description of the topic modeling experiments and their settings

As LDA requires the user to specify the number of topics to be generated, we set this parameter at 20, 50, and 100, generating models for each specification. Using these topic model visualizations, we recorded observations about the accuracy of the lemmatizer and noted the following problems: a number of words that should have been filtered were present in high frequency (examples include καί ("and", κἄν "and", and πῶς "how"); as well, we observed a number of spelling variations. Some authors of the Hellenistic period intentionally wrote according to more traditional orthographical standards, and thus the Hellenistic θάλασσα ("sea") would be spelled as the older, Attic variant, θάλαττα ("sea"). Other examples of this included φυλάττω ("to watch"), and πράττω ("to do").[12] We also observed a number of spelling errors or non-lemmas, such as {, [, ς. Our results at this point indicated that we needed to provide manual corrections to the lemmatizer.

In order to provide these corrections, we used a corpus of several hundred authors from the period 300 BCE—300 CE (24M tokens in 1.3M sentences) in order to gather as much relevant data as possible with our given resources. After processing this corpus, we produced a list of all of the tokens, with their lemma and POS information, ranked by frequency. We then manually corrected the data down to a frequency of 1000 in the automatically annotated corpus. At this point, we had to make the following decisions about how to correct the lemmatizer: how would we classify lemmas that can be tagged as multiple parts of speech at different times (καί ["and"], for example, can plausibly be labeled both ADV or CONJ)? And, how would we annotate tokens that were not words? For the former we chose to retain the multiple POS entries, as most of the words that fell into these categories would be filtered out of the corpus we used for testing anyways. For the latter, we chose to tag those words as NONE, and included them in the POS filtering we applied to the updated corpus. We post-processed the results of the lemmatizer by re-training it with the new, manually edited list and then generating an updated version of the O'Donnell corpus with the superior token/lemma data. The results of this process were then used to generate updated topic models for each of the 20, 50, and 100 topic parameters, as well as to generate a list of semantically related words. All of our topic model tests placed the λ-scale at 0.6 for consistent salience. That is, the topic models represented a balance (between 0.1 and 1.0) of most frequent (1.0 on the λ-scale) to most unique tokens (0.1 on the λ-scale). This setting is visible at the top-right right corner of Figure 1.

[11] pyLDAvis is a Python port of the R package LDAvis (https://github.com/bmabey/pyLDAvis). For more on LDAvis, see [9].

[12] It is important to note that no synchronic variant of Greek is 'pure', and thus the lemmatizer can be intentionally aimed at a particular synchronic period in order to provide a better statistical foundation relevant to that period—in this case, the Hellenistic period.

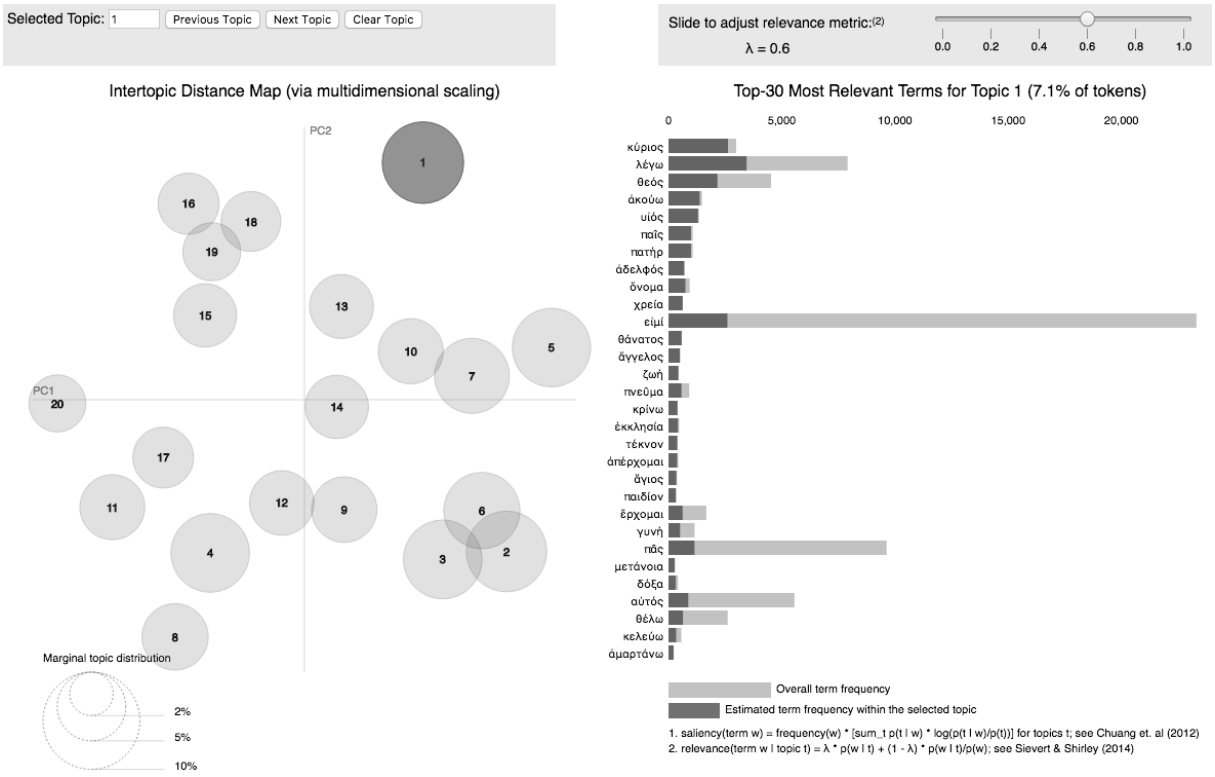

Figure 1. Example of the updated 20-topic model.

5 Measures of Improvement

We tested for three kinds of improvement with the topic models: 1) we checked for a decrease in the number of erroneous lemmas highlighted by the topic models, and 2) we tracked a number of terms that can be described as "familial" terms (including μήτηρ—"mother", πατήρ—"father", υἱός—"son", θυγάτηρ—"daughter", ἀδελφός—"brother", ἀδελφή—"sister", τέκνον—"child", and παῖς—"child"). While a decrease in errors can be empirically measured and depends on the amount of manual correction and post-processing applied to the data, an increase in topical coherence is more subjective. However, an increase in coherence is expected when the number of errors decreases, as erroneous data negatively impacts the statistical foundation upon which the model is based, and effects the automatically assessed distribution of every other lemma in the corpus. Because it is difficult to objectively assess the coherence of the topics—we merely track them for the sake of illustrating the changes that took place—we included a third test as well.

In order to provide a secondary, more objective means of evaluating the improvements to the lemmatizer, we also used GenSim's Word2Vec [3] implementation to 3) generate the top ten terms most related to two of the familial terms that were tracked throughout the topic models, υἱός ("son") and ἀδελφός ("brother"). We compared the lists of related terms both from the original corpus and from the updated corpus using two metrics. The first metric was whether or not there were any errors in the lists of similar terms. For the second metric we compared the original and updated lists to the semantic domain 10 "Kinship Terms" in Louw and Nida's lexicon based on semantic domains [2]. We assumed that there would be a correlation between post-processing the output of the lemmatizer and increased similarity between the list of related terms we generated and Louw and Nida's semantic domain 10. As mentioned above, the actual distributional semantics of the topic models are only of secondary importance; we were primarily interested in the quality of the lemmatizer and the improvement thereof that is reflected in the decreased number of high frequency errors.

6 Extrinsic evaluation of the generated models and lists

6.1 Number of errors

The following table lists the number of errors observed in a random sample of five topics from each model. We found that the number of observed errors decreased in each case between the original and updated corpora. In the case of the 100-topic models, the number of errors would be expected to be higher, as more frequent terms were more widely distributed over the topics, which allowed the less frequent terms—which were less likely to have been corrected in the updated corpus—to appear in the topics. In a topic model, an error counts as an erroneous lexeme, which indicates an incorrect token–lemma pairing. These errors were manually corrected in the updated lemmatizer and updated corpus.

	Original Corpus	**Updated Corpus**
20 Topics	4	1
50 Topics	12	1
100 Topics	5	4

Table 2. Error counts in randomly sampled topics.

6.2 Topical distribution of "familial" terms

When tracking the topics distribution of familial terms, we found that the terms tended to be more distributed before our manual correction of the lemmatizer. Greater distribution results when the LDA model identified less semantic relatedness between the words on the basis of their distribution. However, there are many factors that influence the results of topic modeling, and the objective evaluation of word space models in general has been shown to be difficult [8].

In the original 20-topic model, Topic 9 contained πατήρ ("father") and μήτηρ ("mother"), while Topic 17 contained πατήρ. Topic 8 contained τέκνον ("child"), θυγάτηρ ("daughter"), υἱός ("son"), παῖς ("child"), ἀδελφός ("brother"), and πατήρ ("father"). In the updated 20-topic model, Topic 1 contained τέκνον ("child"), υἱός ("son"), παῖς ("child"), ἀδελφός ("brother"), and πατήρ ("father"). Thus the distribution of these terms decreased, indicating that the updated lemmatizer enabled the LDA model to identify greater semantic similarity between these terms in the updated 20-topic model.

For the original 50-topic model, all the familial terms occurring in the top 30 words of the topics are distributed across 5 topics, while in the updated model the terms coalesce into a single topic, Topic 10—with the exception of ἀδελφή ("sister"), which occurs in Topic 29. Thus the updated 50-topic model also generated these familial terms with less topical distribution.

For the 100-topic model, both the original and the updated models distribute the familial terms we tracked in 5 different topics. Therefore, there was no significant difference in the distribution of familial terms before and after the manual corrections to the lemmatizer.

In summary, tracking the topical distribution of familial terms, while not disclosing all of the causal factors at play, does demonstrate that familial terms were more closely grouped together after the manual corrections to the lemmatizer. Tracking these terms is by no means an objective metric for the performance of the lemmatizer, but it is heuristically useful to note that these words, which have been deemed semantically related by other measures (see next section) exhibit a closer semantic relationship in the topic models generated from the updated corpus.[13] Thus, our topic modelling experiments demonstrated the benefit for certain implementations of lemmatizing and post-processing a corpus using tools specified for the Hellenistic Greek variant.

6.3 Original and updated lists of related terms

The following table contrasts the related terms identified for υἱός ("son"), as well as the cosine value for each term (where 1 represents high similarity and 0 represents low similarity):

[13] For discussion as to the nature of the semantic information conveyed in word space models, see [8].

Original Corpus	Updated Corpus
(‘θυγάτηρ’ [“daughter”], 0.9536596536636353), (‘ἀδελφός’ [“brother”], 0.9501139521598816), (‘μήτηρ’ [“mother”], 0.9479004144668579), (‘παῖς’ [“child”], 0.9425955414772034), (‘γεννάω’ [“to birth”], 0.923077404499054), (‘πατήρ’ [“father”], 0.9012559652328491), (‘οἶκος’ [“house(hold)”], 0.897624671459198), (‘ισραηλ [“Israel”]’, 0.8895389437675476), (‘ζεύς’ [“Zeus”], 0.882246732711792), (‘διαδέξαμαι’ [“to succeed”], 0.8770243525505066)]	(‘θυγάτηρ’ [“daughter”], 0.9703212976455688), (‘παῖς’ [“child”], 0.953923761844635), (‘ἀδελφός’ [“brother”], 0.9433830976486206), (‘μήτηρ’ [“mother”], 0.9313665628433228), (‘γεννάω’ [“to birth”], 0.9292192459106445), (‘ζεύς’ [“Zeus”], 0.9127056002616882), (‘πατήρ’ [“father”], 0.8941320180892944), (‘ἰσραηλ’ [“Israel”], 0.8825995922088623), (‘ιωσεδεκ’ [“Josedek”], 0.8731353282928467), (‘ἀδελφή’ [“sister”], 0.8686784505844116)
Number of terms in semantic domain 10: (5)	Number of terms in semantic domain 10: (5)

Table 3. Top-ten semantically related terms for υἱός (“son”).

It should be noted that the final word in Column 1 is an instance of an erroneous lemma; διαδέξαμαι (should be διαδέχομαι (“succeed”). The term ισραηλ (“Israel”) should also be considered an error, as it has no accent or breathing mark. While this term is partially corrected in the updated corpus (still unaccented), the updated list includes another term with the same problem (ιωσεδεκ [“Josedek”], which like ισραηλ [“Israel”] is a proper noun. Note that proper nouns are unaccented in the LXX, the Greek translation of the Hebrew Scriptures from which several of O’Donnell’s texts were selected). Both corpora included the same number of terms from Louw and Nida’s semantic domain “Kinship Terms”, and so the related terms list reflects only minor improvement after manual correction.

This next table contrasts the related terms identified for ἀδελφός (“brother”):

Original Corpus	Updated Corpus
(‘μήτηρ’ [“mother”], 0.9788503050804138), (‘παῖς’ [“child”], 0.9664503335952759), (‘θυγάτηρ’ [“daughter”], 0.9627480506896973), (‘πατήρ’ [“father”], 0.950951337814331), (‘υἱός’ [“son”], 0.9501139521598816), (‘γεννάω’ [“to birth”], 0.9355642199516296), (‘διαδέξαμαι’ [“to succeed”], 0.9302232265472412), (‘οἶκος’ [“house(hold)”], 0.9233746528625488), (‘ἀδελφή’ [“sister”], 0.9143659472465515), (‘βασιλεία’ [“kingdom”], 0.902644157409668)	(‘μήτηρ’ [“mother”], 0.9736356735229492), (‘πατήρ’ [“father”], 0.9681220054626465), (‘παῖς’ [“child”], 0.9608191251754761), (‘θυγάτηρ’ [“daughter”], 0.9569653272628784), (‘υἱός’ [“son”], 0.9433830380439758), (‘ἀδελφή’ [“sister”], 0.9302892088890076), (‘γεννάω’ [“to birth”], 0.9278334379196167), (‘ἰσραηλ’ [“Israel”], 0.9109088182449341), (‘βασιλεία’ [“kingdom”], 0.907575249671936), (‘ζεύς’ [“Zeus”], 0.9070873856544495)
Number of terms in semantic domain 10: (6)	Number of terms in semantic domain 10: (6)

Table 4. Top-ten semantically related terms for ἀδελφός (“brother”).

The results for ἀδελφός (“brother”) are largely the same as those of υἱός (“son”): there is no significant improvement apart from the correction of the erroneous lemma διαδέξαμαι (“to succeed”), and the introduction of the misspelled term ἰσραηλ (“Israel”). One of the reasons for the misspelling of this term could be either oversight in our manual correction, or else the presence of multiple instances of the token/lemma ἰσραηλ (“Israel”), with different spellings. The presence of multiple, variously spelled instances of token/lemma pairs was one of the most common problems in the initial token/lemma list. Thus for this example as well, the related terms list reflects only minor improvement after manual correction.

6.4 Comparison with Ancient Greek lemmatizer

In order to provide a limited control on our test, we compared the results of using an annotated corpus of Ancient Greek, the AGDT. Using lemmatizer and POS tagger models trained from this corpus as well

as the SBLGNT, we used both the Ancient and Hellenistic Greek models to lemmatize a short sample of 100 lines selected from three different works in the test corpus, the Didache (CE 2; 2241 words), the Shepherd of Hermas (CE 2; 27819 words), and the epistle of Ignatius of Antioch to the Ephesians (CE 1-2; 7956 words).[14] In order to measure the relative precision of each model, we compiled a manually corrected list of tokens, lemmas, and POS tags, and then compared this standard against the results of both models. The results are compiled in Table 5:

	Hellenistic Greek Model		**Ancient Greek Model**	
Part of speech errors	55 / 320	17.2%	65 / 320	20.3%
Lemma errors	58 / 320	18.1%	58 / 320	18.1%
Total errors	**113 / 640**	**17.7%**	**123 / 640**	**19.2%**

Table 5. Comparison of Ancient and Hellenistic Greek models.

The number of tokens in the 100 input lines totaled 320, which created the potential for up to 640 errors that each model could have made. What this small sample set showed us is that, while both models performed similarly, the Hellenistic model worked 1.5% better. However, almost all of the recorded errors involved unique tokens. When the tokens with a frequency of 1 are removed from consideration, and the frequency of the erroneous terms is factored in, the error counts take on a slightly different significance, as seen in Table 6:

	Hellenistic Greek Model		**Ancient Greek Model**	
Part of speech errors	4 / 119	3.4%	15 / 119	12.6%
Lemma errors	4 / 119	3.4%	4 / 119	3.4%
Total errors	**8 / 238**	**3.4%**	**19 / 238**	**8.0%**

Table 6. Error counts when factoring in token frequency.

In light of this recalculation, the Hellenistic Greek model resulted in a lemmatized test sample with 4.6% less errors than the Ancient Greek model. In terms of the larger test corpus, this is a not-inconsiderable difference in performance if distributional word space modeling such as topic modeling is the goal. Keep in mind that this comparison does not take into account the benefit of post-processing. While post-processing would improve the performance of either model, the amount of post-processing is significantly reduced when using a training set of language that more closely approximates the target language variety.

7 Conclusions and future work

In summary, our paper describes the process of tailoring a basic processing tool including a POS tagger and a lemmatizer to operate within the domain of Hellenistic Greek, as well as the task of fine-tuning the tool on the basis of observed errors in the generated data and models. In order to test and exemplify the improvements made through manual correction of the lemmatized data, we used the data to create topic models and lists of semantically related terms. The results were mixed, but we successfully demonstrated the importance of manually tailoring LT tools, by, for example, normalization of orthographic differences among different authors, to better convey information relevant to the period under scrutiny. From our analysis, we can project that more post-processing will improve the lemmatizer. Analogously, a more finely tuned corpus would improve the resulting data—though it would not have an effect on the lemmatizer itself. More manually annotated texts would serve to improve training corpora and, as a result, the tools trained on them. Future research should focus on the effects of combining training data

[14] Thanks to our anonymous reviewers who suggested this further comparison. These three works were selected as a sample because they do not overlap with either of the training sets, but fit within the CE 0-199 timeframe.

on the basis of external criteria such as formality and genre, rather than simply date, and selectively applying specific lemmatizer models to subsets of a corpus.

The main contribution of this paper is its illustration of the importance of targeting machine learning tools toward specific datasets. Through attempting to target Hellenistic Greek, we identified errors and issues for lemmatizing Hellenistic Greek texts, provided evidence that annotations of Ancient Greek texts is less adequate for model training than the Greek New Testament, and provided an initial foray into the use of word space tools in this area of research.

References

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, volume 3, pages 993–1022.

[2] Louw, Johannes P. and Eugene A. Nida. (Eds.). 1988. *Greek-English Lexicon of the New Testament Based on Semantic Domains* (Vols. 1–2). New York: United Bible Society.

[3] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR, 2013*, pages 1–12, Scottsdale, AZ.

[4] Müller, Thomas, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle.

[5] O'Donnell, Matthew B. 2005. *Corpus Linguistics and the Greek of the New Testament*, pages 136, 164–65, Sheffield: Sheffield Phoenix.

[6] Pang, Francis G. H. 2016. *Revisiting Aspect and Aktionsart: a Corpus Approach to Koine Greek Event Typology*, pages 6-35, Leiden: Brill.

[7] Řehůřek, Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta: ELRA.

[8] Sahlgren, Magnus. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm.

[9] Sievert, Carson and Kenneth E. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, MD: Association for Computational Linguistics.