Creating a richly annotated corpus of papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language

Alek Keersmaekers
KU Leuven, Belgium

Abstract

This article describes a first attempt to annotate the full Greek papyrus corpus automatically for linguistic information. It gives an overview of existing work on Ancient Greek and analyzes the typical problems one encounters when using natural language processing techniques on (1) a historical corpus of (2) a highly inflectional language (as opposed to the more analytic present-day English) and offers solutions to them, testing several different approaches. The focus is on part-of-speech/morphological tagging and lemmatization; some syntactic parsing experiments are also briefly discussed. The conclusion discusses the strengths and shortcomings of the examined techniques and suggests possible ways to further improve tagging and parsing accuracy.

Correspondence:
Alek Keersmaekers,
Blijde-Inkomststraat
21 - box 3308, 3000 Leuven,
Belgium.
E-mail:
alek.keersmaekers
@kuleuven.be

1 Introduction

The Greek documentary 'papyri',[1] a corpus containing about 4.5 million words of non-literary texts ranging from the 3rd century BC to the 8th century AD, is an invaluable source for the study of the history of the Greek language. Not only does this body of texts contain the largest amount of non-literary Ancient Greek material that is available to us but their writers are also sociolinguistically far more diverse than the (typically male, elite) authors of literary Greek texts.

The number of studies focusing on the language of these texts is rather restricted. This can partly be explained by the lack of a linguistically annotated corpus, making searching the papyri far from obvious. Hence, this article will describe a first attempt to come to a full automatic annotation of morphology and lemmata in all papyri (as well as briefly discussing some syntactic parsing experiments).[2] Due to the fact that Greek is a highly inflected language, and due to the fragmentary transmission of the papyri, these texts pose several challenges for natural language processing. The aim of this article is therefore to present an overview of the problems one encounters when trying to annotate these texts automatically (Section 2) and explain the performance of different competing techniques in this context (Section 3), after briefly discussing existing research on natural language processing (NLP) techniques for Ancient Greek (Section 1). Finally, it will summarize the main findings and offer some perspectives for future research (Section 4).

2 State of the Art

2.1 Available material

XML versions of the papyrus texts have been made available publicly by the Duke Databank of

Digital Scholarship in the Humanities © The Author(s) 2019. Published by Oxford University Press on behalf of EADH.

doi:10.1093/llc/fqz004

Documentary Papyri (Cayless *et al.*, 2016). Alongside with the 'raw' texts, the XML files also include tags indicating editorial regularizations (spelling corrections, interpretations for missing text, etc.). As for linguistically annotated papyrus texts, a first attempt to annotate the papyri manually was undertaken by Porter and O'Donnell (2010), who tagged forty-five papyrus letters for morphology, syntax, and sociolinguistic and pragmatic information. Their corpus has not yet been publicly released. A more comprehensive effort has been undertaken by the Sematia project, described in Vierros and Henriksson (2017). They have built a tool to tokenize the papyrus texts and are in the process of making manually annotated dependency treebanks of (a subset of) the corpus through the help of the annotation platform Arethusa (a tool developed by the Perseus project to annotate Ancient Greek and Latin texts). Currently 115 annotated texts have been released on https://sematia.hum.helsinki.fi.

Aside from the papyri, several manually annotated (literary) Ancient Greek texts are publicly available. There are currently two major dependency treebanks: Perseus' AGDT (Ancient Greek Dependency Treebanks, Bamman and Crane, 2011; 549,906 tokens as for the 2.1 release), containing Archaic, Classical, and Post-Classical poetry and prose annotated for lemma and morphological and syntactic (and in a few cases semantic) information; and the PROIEL treebanks (Pragmatic Resources in Old Indo-European Languages, Haug and Jøhndal, 2008; 247,726 tokens), containing editions of the Greek New Testament, parts of Herodotus' Histories and Sphrantzes' Chronicles, annotated for lemma and morphological, syntactic, and pragmatic information. Both treebanks are also released in the Universal Dependencies project (Nivre *et al.*, 2016, see http://www.universaldependencies.org). In addition, there are some isolated projects offering a number of morphologically annotated texts (see Section 3.2).

As for the sociohistorical background of the papyri, the Trismegistos project—a set of databases containing historical 'metadata' for all papyrus texts—covers this information most extensively.[3] These databases not only include general information about each papyrus text (e.g. date, place, and genre information)[4] but also more specific information about certain entities or words occurring in the texts (e.g. personal names, place names, and editorial and scribal text regularizations). Section 3 will describe how these data were used in NLP approaches to these texts.[5]

2.2 Automated approaches

There have been several attempts already to process Ancient Greek morphologically and syntactically. A morphological analysis tool of Greek, called Morpheus, has been developed by Crane (1991). It generates all possible lemmas and morphological analyses—i.e. inflectional information such as case, gender, and tense—for a given Ancient Greek word form and can cope with most dialectal and historical variation. An open-source version is publicly available,[6] to which missing lemmas and word endings can be easily added (see also Section 3).

Some scholars have explored stochastic approaches to morphology/part-of-speech tagging and syntactic parsing of Ancient Greek. Dik and Whaling (2008) made use of TreeTagger (Schmid, 1994), supplied with a lexicon generated by Morpheus, to tag literary classical Greek texts automatically. They reported an accuracy of about 91% when tested on a sample of 2,000 words of the rhetor Lysias. Lee *et al.* (2011) compared the performance of a standard part-of-speech tagging model to a joint morphological/syntax model for several inflectional languages, including Ancient Greek. They found that joint models slightly improve morphological tagging accuracy for all morphological attributes, as well as syntactic parsing accuracy. Mambrini and Passarotti (2012) tested the accuracy of the dependency parser MaltParser (Nivre *et al.*, 2007) trained and tested on poetry from the AGDT (using gold morphology and lemmata). Their highest result was a labeled attachment score (Labelled Attachment Score (LAS), percentage of tokens for which both the syntactic head and the dependency relation are correctly identified) of 71.72%, when the parser was trained and tested on Homeric Greek. Recently, Celano *et al.* (2016) have compared several part-of-speech taggers: Mate, Hunpos, RFTagger, OpenNLP, and NLTK Unigram

Tagger. Mate gave the best results: 88% accuracy when trained and tested on the data from the Ancient Greek Dependency Treebanks. They argued that most remaining errors can be explained by inconsistencies in the training data.

3 Problems

English has so far been the language that has received most attention in research on natural language processing. Yet due to both linguistic and genre differences between English and (Ancient) Greek, techniques that are successfully applied to English texts do not guarantee the same level of performance if applied to Greek papyri. This section will describe the most prominent problems researchers have encountered when trying to process Greek and other highly inflectional languages, as well as some specific problems regarding the genre and textual transmission of papyrus texts, focusing on morphological analysis.

3.1 Linguistic problems

In contrast to English, Greek conveys more information (e.g. aspect, voice, and alignment) through morphological means, while English would represent the same information analytically. Therefore traditional 'N-Gram'-based approaches, which are quite suitable for English, might encounter problems analyzing Ancient Greek. The following issues are particularly relevant for natural language processing:

(a) Inflectional languages typically have a very extensive tag set. Whereas the Brown English tag set counts no more than 200 tags (see Hajič and Hladká, 1998), the tag set of inflectional languages can amount to several thousands, given that tags indicate, apart from the determination of the part-of-speech category, also multiple morphological categories (e.g. noun + singular, feminine, and dative).[7] As a consequence, the amount of possible outcomes to be considered by a tagger is much higher. It comes as no surprise that this has a considerable impact on tagging accuracy. Several techniques have been proposed to deal with this. One possibility consists in making the tagger, in a first stage, identify the part-of-speech only and gradually identify the other features afterwards (as in Acedański, 2010). Another approach is to determine each morphological attribute (including part-of-speech) individually and calculate tag probability as the product of the probabilities of each morphological attribute (Schmid and Laws, 2008; Sawalha and Atwell, 2010).

(b) As a result of the large tag set, the amount of possible features the part-of-speech tagger may consider is also relatively large.[8] While hidden Markov models (HMMs) are quite popular for English, different machine learning models, such as maximum entropy (Ratnaparkhi, 1996) and conditional random field (Lafferty *et al.*, 2001) models, are typically proposed for highly inflectional languages, since HMMs have difficulties integrating a large amount of features (Adafre, 2005; Ekbal *et al.*, 2008). Another method consists in using decision trees to ensure that the statistically most relevant features in a given tag context will be considered (Schmid, 1994).[9]

(c) Another consequence of the size of the tag set is the large amount of 'unseen' word forms, as the amount of possible word forms is too large (due to inflection) to be fully represented in the training data. Hajič (2000) argues that the best solution for this problem is to analyze the test data first with a language-specific morphologic analysis tool. The tagger can then use this 'dictionary' to look up forms that it has not seen in the training data. Integrating the output of a morphological analyzer often has a considerable positive impact on tagging accuracy for inflectional languages: see e.g. Dik and Whaling (2008) for Ancient Greek (using the morphological analysis tool Morpheus, see Section 1.2); Habash and Rambow (2005) and Denis and Sagot (2009) for other languages.

(d) While the word order of English is quite rigid, most inflectional languages (especially Ancient Greek, see Dik, 2007) have a far more flexible word order. As a result, it is

far from obvious that machine learning approaches that assume a relatively predictable ordering of words (e.g. *N*-gram-based approaches) would have a similar performance for Greek as for English.[10] While there is not a large amount of research on the impact of free word order on part-of-speech tagging, Dik and Whaling (2008) argue that 'a trigram Markov model [is] in fact capable of modeling Greek grammar remarkably well'.[11]

(e) Since some syntactic information such as alignment is expressed at the morphological instead of the syntactic level, morphological and syntactic analysis are strongly interrelated in inflectional languages (Lee *et al.*, 2011). Hence performing morphological and syntactic analysis jointly instead of in a pipeline model (implying that the two tasks can be performed independently) often improves accuracy for both tasks (see e.g. Cohen and Smith, 2007; Bohnet *et al.*, 2013; Lee *et al.* (2011) also note superior results for both tasks with Ancient Greek).

3.2 Textual transmission

While the previous section discussed general linguistic properties of Greek, another set of problems arises due to the (fragmentary) way the papyrus corpus is preserved. While we do always possess an original version of the papyrus—unlike literary texts, which are almost always transmitted to us through subsequent copying by medieval scribes—this original version often contains several gaps due to physical damage to the papyrus. In addition, the spelling is not standardized. In this respect, the papyri have much in common with other genres that are difficult to analyze automatically such as tweets (Gimpel *et al.*, 2011). However, unlike tweets, most papyrus texts have been standardized by modern editors. While a standardized spelling is highly beneficial for NLP tasks, editors often also correct morphosyntactic problems such as case usage, which might lead to misleading results when the data are analyzed automatically: e.g. if one automatically corrects a dative to an accusative due to an editorial regularization, it will also be automatically analyzed as an accusative, although one might want to preserve the very fact that the original text has a dative. On the other hand, for some tasks, e.g. dependency parsing, even grammatical corrections may be beneficial: as the parsers are mostly trained on highly regularized literary Greek, it may be useful to have the test corpus closely align with the training data grammatically as well. In other words, it is necessary to strike a balance between making the test corpus as easy as possible to analyze automatically and still preserving all linguistic information that is present in the data.

Modern editors also often try to supply missing text fragments, for instance on the basis of texts with analogous language use and comparable context. However, at times the papyrus is too damaged to reconstruct the missing text, implying that strategies need to be developed to handle such incomplete sentences. While this might not be such an acute problem for more 'local' tasks such as lemmatization and part-of-speech tagging, it goes without saying that syntactic parsing, which operates at the sentence level, will become far more difficult when one or multiple words are missing.

3.3 Training versus test corpus

The current work on automated linguistic processing and linguistic annotation of Ancient Greek has so far focused on literary Greek. As a result, the available linguistically annotated data (as mentioned in Section 1) to be used in a supervised machine learning approach is largely literary: in total, the training corpus I collected for part-of-speech tagging consists of 971,638 tokens of literary Greek and only 38,539 tokens of documentary papyrus text (see also Section 3.2). There are also considerable chronological differences between the (literary) training data and the papyrus data to be analyzed: the training data include Classical and early Post-Classical Greek texts (5th century BC–3rd century AD), while the test data are only Post-Classical (3th century BC–8th century AD). This is problematic, since tagging accuracy has been shown to decrease markedly when out-of-domain data are used.

One simple solution consists in adding more manually annotated papyrus data to the training corpus: therefore we expect tagging accuracy to improve when more data from the SEMATIA treebanks

(Vierros and Henriksson, 2017) are available. Another method is to integrate information from the unannotated target (papyrus) corpus during part-of-speech tagging: while the corpus is likely too small to do the tagging completely unsupervised (Goldwater and Griffiths, 2007; Das and Petrov, 2011; the unsupervised unsupos tagger described in Biemann, 2006, has been implemented in Java), some domain adaptation methods used for other domains (e.g. biomedical text tagging trained on data from the *Wall Street Journal*) could also be useful for the papyri (Blitzer *et al.*, 2006, Daumé, 2007; see Schnabel and Schütze, 2014, for a practical implementation using word vector representations).[12]

4 Own Approach

This section describes the pipeline model used to process the Greek papyri linguistically, i.e. tokenization, part-of-speech/morphological tagging, and lemmatization. It will also briefly discuss some first syntactic parsing experiments. For each step, I will describe the methods I used to handle the problems mentioned in Section 2.

4.1 Preparatory work: Tokenization and related problems

As the XML versions of the texts contain no markup for individual words, the papyri first needed to be tokenized. This task was relatively easily tackled, since word boundaries can simply be identified by relying on whitespaces and punctuation marks, which are supplied by the editor (the original Greek was written continuously).[13] However, some problems arose due to problems in the XML version of the text. These included missing spaces between words and the capitalization of words other than proper names at the beginning of a sentence (as it is the convention for the papyrus corpus to only capitalize proper names regardless of the position of the word in the sentence). These problems were corrected semi-automatically: in the case of missing spaces, for instance, the morphological analysis tool Morpheus was used to check which possible split of the conjoined word consists of two valid Greek words. Afterward, I checked the output manually.

Each token was assigned an 'original' form (i.e. the form as it is preserved in the text) and a 'regularized' form (i.e. the form corrected by the modern editor). Afterward one of the two versions of the token could be chosen dynamically for each NLP task. For part-of-speech/morphology tagging the regularized version will be chosen when only the spelling is corrected, while the original version will be used in the case of morphological corrections (see Keersmaekers and Depauw, forthcoming, for a more detailed description of the procedure). This way a more 'standard' version could be used when it would have no impact on the morphology, while in other cases the word would be tagged with the morphological attributes that are used in the text.[14] For lemmatization and syntactic parsing, the regularized version of the token was used in all cases, since using the highly irregular original version of the text would likely have a considerable negative impact on the output.[15]

4.2 Part-of-speech and morphological tagging

In a following stage, the papyri were analyzed for part-of-speech and morphology information. I prepared a manually morphologically annotated papyrus corpus of 2,378 tokens[16] of letters and petitions as test data. I tested three part-of-speech taggers—RFTagger (Schmid and Laws, 2008), MarMoT (Müller *et al.*, 2013), and Mate (Bohnet *et al.*, 2013)—which were specifically chosen to tackle the problems mentioned above:

- RFTagger, specifically developed for languages with large tag sets, uses a HMM as its machine learning model. It determines each morphological feature on an individual basis and calculates tag probability by multiplying the probabilities of each individual feature, therefore being able to handle complex tags such as those of Greek well. RFTagger relies on decision trees to select the most relevant contextual features to be used during the tagging, so that the large amount of morphological features of Greek is no hindrance. The tagger can be supplied with an external morphological lexicon. If this is the case, only morphological analyses that are present either in the lexicon or in the training data will be considered for a given word form, unless

the form occurs in neither (in which case the tagger exclusively tries to determine the correct analysis on the basis of the word form and on part-of-speech tag frequencies). In other words, the lexicon is used as a 'hard constraint', as it restricts the amount of possible analyses that will be considered to only a select few.

- MarMoT uses so-called 'pruned' Condition Random Fields—which are well suited for datasets with a large amount of features, see Section 2—that allow for higher-order models (see Müller *et al.*, 2013). Just like RFTagger, MarMoT also decomposes part-of-speech tags into individual morphological attributes. This tagger can also be supplied with a morphological lexicon, although occurrence of the form in the lexicon is simply one of the features in the model. In other words, the lexicon is a 'soft constraint', since analyses that are not present in the lexicon can still be chosen (but are less likely).
- Mate is a joint morphological tagger and syntactic (transition-based) dependency parser—although these two steps can also be executed in a pipeline—using a structured perceptron learning model for tagging (as well as parsing). It splits up part-of-speech tags in the part-of-speech proper and morphological information, while the morphology is still treated as one unit. Like the other two taggers, it can also be supplied with a lexicon, which is treated as a soft constraint like MarMoT.

I have supplied all of the taggers with a morphological lexicon automatically generated by the Ancient Greek morphological analysis tool Morpheus (Crane, 1991). Since Morpheus was originally designed to analyze literary texts, it does not contain some frequent forms in papyrus texts (in particular Latin loan words). Therefore I expanded Morpheus' vocabulary beyond literary Greek by adding the most frequent forms not recognized after a first tagging iteration manually to its lexicon. For all taggers, out-of-the-box settings have been used.[17] They have been trained on the prose of the Ancient Greek Dependency Treebanks (Bamman and Crane, 2011; v. 2.1 release), combined with the MorphGNT analysis of the New Testament (Tauber, 2017) and the CCAT tagging of the Septuagint (Kraft, 1988). Table 1 shows the accuracy of each part-of-speech tagger—i.e. the percentage of analyses that have both part-of-speech and morphological information correct—on the test data.

Table 1 Accuracy of part-of-speech/morphological taggers on papyrus test corpus ($N = 2{,}378$)

	Accuracy
RFTagger	0.947
MarMoT	0.947
Mate	0.909

These figures are higher than those of previous applications of part-of-speech tagging to Ancient Greek—Dik and Whaling (2008) report an accuracy of 91% using TreeTagger, while the maximum accuracy Celano *et al.* (2016) achieve (with Mate) is 88%. As the test corpus is different, however, and a slightly different tag set than the one of Dik and Whaling (2008) and Celano *et al.* (2016) is used,[18] comparing is difficult. Nevertheless, both RFTagger and MarMoT seem to handle the morphological complexity of Greek well. By decomposing part-of-speech tags and (in the former case) using decision trees to select the most relevant contextual features or (in the latter case) feature integration in a conditional random field model, the taggers can deal with the large tag set of our corpus [Problems (a) and (b) described above]. The use of a morphological lexicon (also used by Dik and Whaling, 2008 but not by Celano *et al.*, 2016) is a valuable help for the tagger to cope with 'unseen' word forms [Problem (c)]—without lexicon RFTagger's accuracy dropped 2.4 points, to 92.2%. As described above, RFTagger treats this lexicon as a 'hard' constraint (i.e. only analyses present in the lexicon or training data are considered), while it is a 'soft' constraint with MarMoT. As a consequence, in almost all cases RFTagger generated an analysis which could be a correct morphological interpretation of the word, while MarMoT sometimes generated an analysis that is theoretically impossible: for instance the word ἀρουρῶν (*arourōn*, genitive plural of *arourá* 'field') was once tagged as masculine by MarMoT, even though the only possible analysis of the form is feminine. Most exceptions concerned forms with a wrong accent (added by the editor): the form ταυροῖς (Trismegistos (TM) 11099, l. 7),

for instance, was tagged by RFTagger as the very infrequent verb ταυράω 'to want the bull' instead of the noun ταῦρος 'bull' (which has the accent ταύροις). As Morpheus is accent-sensitive, it did not consider the nominal analysis as an option. Since MarMoT also allows analyses that are not present in the lexicon, however, ταυροῖς was correctly tagged as a noun (unlike with RFTagger). In such cases a less restrictive use of the lexicon can be beneficial (and is probably closer to human language processing); however, such accent errors can also easily be corrected automatically.

Mate's accuracy was far lower than the other two taggers. This could be due to several factors: (1) the tagging model could be unsuitable for Greek, (2) the (smaller) amount of training data could hurt tagging accuracy,[19] or (3) the joint parsing model could be detrimental to the tagging process, possibly due to low parsing accuracy. As for Factor 2, while RFTagger scored a little lower when it was trained on the same training data as Mate (94.1% accuracy instead of 94.6), it was still far above the 90.9 accuracy of Mate. Regarding Factor 3, tagging accuracy was even lower when testing a non-joint tagging model with Mate (90.0 versus the 90.9% accuracy of the joint model). In fact, the joint tagging/parser model was able to tag some syntactic constructions correctly—especially involving long-distance relations—which neither the non-joint Mate model nor RFTagger and MarMoT were able to. Two examples:

In Example 1, the adjective ὀλίγον (*olígon*: 'little') can be either nominative or accusative (as for neuter nouns and adjectives the suffix *-ον* is a homonym in both cases). From the use of the copula ἔστι (*ésti*: 'is'), however, we know that it should be nominative, as it is used as a predicative adjective. An *N*-gram model could only theoretically pick up this information if *N* is extended to 8, while the more sophisticated syntactic model that Mate uses gave the correct analysis. Likewise, in Example 2, the suffix *–ουσι* of the form **πλησιάζουσι** (*plēsiázousi*) can either point toward a dative plural participle ('being near to') or a third-person indicative verb ('they are near to'). As the latter use of *–ουσι* is much more common, it is no surprise that RFTagger and MarMoT tagged it as an indicative verb. Yet from the syntactic analysis of the sentence, we know that the main indicative verb is ἀξιοῦμεν (*aksioũmen*: 'we ask'), so that the correct analysis of **πλησιάζουσι** is instead a participle agreeing with the dative noun **τόποις** (tópois: 'places'). Again, this information is too sophisticated to be picked up by an *N*-gram model, while Mate's syntactic model could handle it correctly. However, these examples are rather rare and even in constructions in which the syntactic structure is often crucial (e.g. confusion between accusative and nominative), Mate performs only marginally better or worse than the other taggers.[20]

(1)	**ἔστι**	γὰρ	τὸ	**πλῆθος**	τοῦ
	ésti	gár	tó	plēthos	toũ
	be.3.sg	since	the.NEUT.NOM.sg	quantity.NEUT.NOM.sg	the.NEUT.GEN.sg
	ἀργυρίου	**οὐκ**	**ὀλίγον**		
	arguríou	ouk	**olígon**		
	money.NEUT.GEN.sg	not	**little.NEUT.NOM.sg**		

'Since the quantity of the money isn't small.' (TM 5364, l. 6)

(2)	ἀξιοῦμεν	(10 more words)	εἰς	τὸ	δύνασθαι	ἡμᾶς	**ἐν**
	aksioũmen		eis	tó	dúnasthai	hēmãs	en
	ask.1.pl		for	the.NEUT.ACC.sg	be.able.to.INF	we.ACC	in
	τοῖς	**συνήθεσι**		**τόποις**	**ἐργαζομένους**		
	toīs	sunḗthesi		tópois	ergazoménous		
	the.MASC.DAT.pl	usual.MASC.DAT.pl		place.MASC.DAT.pl	work.PART.MASC.ACC.pl		
	πλησιάζουσι			τῇ	κώμῃ		
	plēsiázousi			tẽi	kṓmēi		
	be.near.PART.MASC.DAT.pl			the.FEM.DAT.sg	village.FEM.DAT.sg		

'We ask (. . .) so that we be able to work in the usual places that are near to the village.' (TM 14145, l. 15-20)

In sum, while joint morphological and syntactic analysis seems to have similar potential for Greek as for other inflectional languages, Mate's low accuracy seems to be primarily caused by its tagging model that is unsuitable to analyze Ancient Greek. A major difference between *Mate* and the other two taggers is the way it treats morphological descriptions: while Mate would treat, e.g. singular + masculine + dative as one unit, RFTagger and MarMoT determine each morphological attribute individually. Presumably this is an important contributing factor why RFTagger and MarMoT perform better than Mate, since the morphology of Greek might be too complex to treat as an atomic unit. Moreover, Mate also seems to have more difficulties than the other two taggers to integrate lexical knowledge in its model, as several words received an analysis that was neither present in the training data nor in the lexicon: e.g. in TM 961, l. 6, **ποιήσαις** was analyzed as a future indicative, even though the only form present in the lexicon was an aorist optative.

Table 2 shows tagging accuracy for each individual morphological attribute.[21]

Gender, mood, person, and case are consistently the most difficult features to determine. This is not surprising, since these categories contain several ambiguous forms that the taggers struggled with—mainly confusion between masculine and neuter in most case forms of adjectives on -ος (e.g. δικαίου *dikaíou*, genitive masculine/neuter singular of δίκαιος *díkaios*) and between feminine, masculine, and neuter plural (e.g. **αὐτῶν** *autôn*, genitive masculine/feminine/neuter singular of **αὐτός** *autós*), between indicative and subjunctive in forms ending in –ω (e.g. **παρενοχλῶ** *parenokhlô*, subjunctive or indicative first-person singular of **παρενοχλέω** *parenokhléō*), between first-person singular and third-person plural imperfect and some aorist forms (e.g. **ἔσχον** *éskhon*, first-person singular or third-person plural aorist of **ἔχω** *ékhō*), and between nominative and accusative of neuter nouns (e.g. **ἔργα** *érga*, nominative or accusative plural of **ἔργον** *érgon*). Most of these features (except for person when there is only one verb in the sentence) can be determined accurately when the syntactic function of the word in the clause in known, again suggesting that joint syntactic and morphological analysis could solve most remaining errors.

Table 2 Tagging accuracy by morphological attribute

	Median	RFTagger	MarMoT	Mate
Derivative category	0.996	0.996	0.996	1.000
Part-of-speech	0.994	0.994	0.994	0.972
Number	0.990	0.990	0.995	0.979
Voice	0.989	0.989	0.993	0.965
Tense	0.985	0.985	0.987	0.919
Degree	0.974	0.974	0.987	0.765
Case	0.974	0.976	0.974	0.960
Person	0.963	0.963	0.977	0.949
Mood	0.959	0.959	0.977	0.894
Gender	0.951	0.953	0.951	0.933

Table 3 Sub-parts of the training data and their effect on tagging accuracy

Corpus	Tokens	Tagging accuracy (relative)
1) Adding data		
Aeschylus	46,745 (4.6%)	−0.23%
Hesiod	18,866 (1.9%)	−0.23%
Sophocles	48,644 (4.7%)	−0.28%
Homer	232,336 (19.2%)	−0.51%
2) Removing data		
Aesop	5,166 (0.5%)	+0.14%
Plato	6,086 (0.6%)	+0.05%
Apollodorus	1,229 (0.1%)	−0.05%
Diodorus	25,528 (2.6%)	−0.05%
Plutarch	21,870 (2.2%)	−0.09%
Lysias	7,123 (0.7%)	−0.23%
Athenaeus	44,741 (4.6%)	−0.28%
Septuagint	654,322 (66.9%)	−0.28%
Papyri	5,788 (0.6%)	−0.32%
New Testament	152,772 (15.6%)	−0.37%
Thucydides	24,901 (2.5%)	−0.46%
Polybius	28,080 (2.9%)	−0.51%

To test whether the mismatch between training data and test data had a significant effect on tagging accuracy (i.e. the training data mostly contained literary prose, while the test data was non-literary), I checked the effect of (1) adding poetic data to the test corpus, which is even further removed stylistically from the test corpus and (2) removing several prose authors from the test corpus, using RFTagger. Table 3 shows the result.

Several findings can be retrieved from the above data. First of all, adding poetic data is clearly detrimental to the tagging process: while removing prose

authors from the training data in most cases has a negative effect on tagging accuracy, the tagger performs better if poetic authors are excluded. Moreover, there is not a clear relationship between the amount of tokens of the subcorpus included in the training data, and the relative impact excluding it from or including it in the training data has on tagging accuracy: while 67% of the training data is from the Septuagint, for instance, excluding it only has a tiny effect on tagging accuracy (−0.3%), while the effect of excluding data from Polybius, which is only 3% of the training data, is even larger (−0.5%). In fact, the papyrological data, which is less than 1% of the training data, has a larger relative impact on tagging accuracy when excluded than most other subcorpora. In this context, it is not surprising that prose authors such as Lysias (who wrote in a relatively unadorned style) and Polybius (a post-classical history writer) have a large positive impact on tagging accuracy relative to their token count, and that Homer, who is stylistically and diachronically the furthest removed from the papyrus corpus, has a considerable negative impact when added to the training data (especially since including him would mean that roughly one fifth of the training data would be Homeric). In other words, the quality of the training data, i.e. the degree to which it resembles the papyrus corpus, seems to be far more important than its quantity.[22]

I also briefly investigated the effect of missing words due to physical damage to the papyrus. For sentences with one or more words missing (which could not be supplied by the editor), tagging accuracy with RFTagger was 0.954 (829/869 words tagged correctly), while it was 0.942 (1,422/1,509 words) for 'complete' sentences. In other words, missing words clearly have no negative effect on tagging accuracy, likely due to the short-context model (3-grams) that was used.

4.3 Lemmatization

In a following stage, the papyrus data were lemmatized. Due to the scarcity of trainable lemmatizers (in comparison to part-of-speech taggers), I only tested Lemming (Müller *et al.*, 2015), a lemmatizer developed together with MarMoT. Lemming is trained on a morphologically annotated text corpus and uses formal features, lemma frequencies and part-of-speech/morphological information. It can be supplied with several resources including a lexicon and lexical cluster data—for the time being I only used a lexicon, i.e. all lemmas included in the *Liddell-Scott-Jones Greek-English Lexicon* (Liddell and Scott, 1940). It can be run jointly together with MarMoT or in a pipeline (in the latter case the part-of-speech information needs to be supplied).

I tested Lemming on a smaller subset of the data used for the part-of-speech tagging task (1,167 lemmas in total) in a pipeline with RFTagger, which had the most accurate result overall. The initial accuracy of Lemming was 0.969, i.e. 1,131/1,167 lemmas were correctly identified. Most errors were due to the complex morphology of Greek, particularly with verbal stem changes: e.g. the passive participle ἐπενεχθεῖσαν (*epenekhtheîsan*) from the verb ἐπιφέρω (*epiphérō*) was identified as the fictive verb ἐπενέκω (*epenékō*), which would be closer to the inflected form formally. Therefore I decided to integrate the morphological analyses of Morpheus within this task as well. More precisely, I modified the code of Lemming so that for forms recognized by Morpheus the lemmatizer only considers lemmas with the same morphological tag—i.e. a hard constraint, since hard constraints also proved to be useful during part-of-speech tagging (see above). When this step was included, the accuracy of the lemmatization task rose from 0.969 to 0.985 (1,150/1,167 lemmas correct). This high accuracy is not really surprising, since Greek encodes much morphological information in its suffixes, so that for most words only a single lemma is possible.

The remaining errors in our test data were mostly cases in which an incorrect lemma was caused by an incorrect part-of-speech tag. An example is the lemma of the form θελήσῃ (*thelésēi*) (TM 36197, l. 6) which was identified as the noun θέλησις (*thélēsis* 'want') because of an automatically generated part-of-speech tag 'noun'. Although this is morphologically possible, the correct analysis in this particular context is the verb θέλω (*thélō* 'to want'). Therefore it might be useful to remove an exact match with the part-of-speech tag from the requirements of our modified version of Lemming

(i.e. include all lemmas generated by Morpheus, regardless of their part-of-speech tag). Possibly calculating part-of-speech and lemma information jointly (which is possible with MarMoT) could also resolve these errors and improve the accuracy of both tasks.

4.4 Syntactic parsing: Preliminary experiments

Finally, I also conducted some syntactic parsing experiments. I trained MaltParser (Nivre *et al.*, 2007) on data from the AGDT (all prose authors), the PROIEL project (the New Testament and Herodotus data), and the Sematia treebanks. MaltOptimizer (Ballesteros and Nivre, 2012) was used to select the most optimal features to parse Ancient Greek, to handle the large amount of possible features that the parser could consider. Using a subset of the test data for the part-of-speech tagging task—1,521 tokens in total, consisting of all the texts that are (almost) completely preserved—the highest LAS I was able to obtain was 0.674, i.e. roughly two-thirds of the tokens had their syntactic head and dependency label correctly identified. Needless to say, there is still much room for improvement. Some of the most difficult structures for the parser to handle were coordination structures, which only achieved at max an LAS of 0.385. Since about 10.5% of the tokens in the test data were involved in a coordination structure (161 in total), finding better ways to parse these structures would also obviously have a considerable impact on parsing accuracy.[23]

5 Conclusions

The goal of this study was to identify the most prominent problems with NLP approaches to the Ancient Greek papyrus corpus—a highly inflectional and historical language—and put forward possible solutions. As for the Greek language, I have identified five main problems concerning the inflectional status of the language in Section 2.1. In Section 3, I showed which tagging approaches can

(a) deal with the large amount of tags that the tagger can consider;
(b) handle the similarly large amount of features that can be integrated in the tagging model; and
(c) interpret the large amount of 'unknown' word forms that do not occur in the training data.

As for (a), splitting up complex morphological tags in the product of the probabilities of each morphological attribute seems to be the best possible way to handle large tag sets such as for Ancient Greek. The two best scoring taggers used a different method to deal with Problem (b)—RFTagger used decision trees to select the most relevant features from the word context, while MarMoT used conditional random fields which are suitable to handle a large amount of features—but both methods proved to be suitable to analyze Greek papyri. As many inflected forms will by nature not occur in the training data, enriching the tagger model with the output of a morphological analyzer seems to be the best possible way to deal with Problem (c), as Hajič (2000) has argued—the same is true for lemmatization, since integrating the output of Morpheus in Lemming was clearly beneficial for the process. This article discussed whether such a lexicon should function as a 'hard constraint', i.e. the tagger should only consider forms that appear in the lexicon, or a 'soft constraint', i.e. the probability of tags should increase when the form appears in the lexicon, but tags that do not appear in it could also be considered. Both approaches have advantages and disadvantages. The first approach strongly constrains the possible search space for the tagger but could be too strict when certain analyses of a word are not recognized by the morphological analyzer, whether due to, e.g., spelling errors or because the analyzer does not completely cover the target language. The second approach, on the other hand, is more lenient in such cases but might also suggest analyses which are theoretically impossible.

I also mentioned the relatively free word order of Ancient Greek as a potential problem in Section 2.1 [Problem (d)]. This problem did not get much attention in this article, since previous approaches to morphological tagging in Ancient Greek did not show the word order of Greek to be particularly problematic, and the remaining problems I found during the automated tagging of the papyrus text corpus also did not seem to be particularly related to word order.[24] However, for other NLP tasks, e.g. syntactic parsing, this problem becomes more

prominent, and specialized approaches are likely needed. More important for morphological tagging is the interdependence of morphology and syntax [Problem (e)]. Almost all remaining part-of-speech/morphological tagging errors indeed are due to complex syntactic relations which are difficult or even impossible to identify by a tagging model that only uses the local context of a word. This strongly suggests that joint morphological and syntactic analysis could break the ceiling that the current pipeline model seems to have reached. However, a suitable tagging model to analyze Ancient Greek as well as a high-scoring parsing model is obviously a necessary prerequisite, as the low accuracy of the Mate tagger/parser shows.

The documentary papyrus corpus in itself also has some particular problems mentioned in Sections 2.2 (spelling variation and uncomplete preservation of the texts) and 2.3 (the lack of annotated papyrus text to train a parser on). First of all, while there is a large amount of spelling (and sometimes morphological) variation, it is possible to regularize the language of the papyrus texts to a large extent due to editorial practices. However, because editors sometimes go too far in regularizing the text (e.g. by changing morphology or syntax as well), caution is needed. By keeping both the 'original' and 'regularized' version of the text, it is possible to choose dynamically which version of a word is preferred for each NLP task. Regarding physical damage to the papyrus, it was shown that physical damage had no negative effect on tagging accuracy, due to the local context that is used for tagging (so that most words do not fall in the scope of such 'gaps'). For syntactic parsing, which takes the structure of the whole sentence into account, this is obviously a more serious problem, and specific strategies (e.g. parsing of partial structures) need to be developed for this task.

As for the nature of the test corpus, a particular difficulty to analyze the documentary papyri was the 'mismatch' between training data and test data, i.e. the former is mostly literary and situated earlier in time, while the latter is non-literary and situated later in time. To cope with this, I added some papyrus data to the training data, which even though it is relatively limited still has a positive impact on tagging accuracy (see Table 3). I also expanded Morpheus' vocabulary beyond literary Greek by adding the most frequent forms not recognized in a first iteration manually to the lexicon (e.g. I added the lemma ποταμοφυλακίς to Morpheus, based on the occurrence of forms such as ποτομαφυλακίδος, ποτομαφυλακίδων, ποταμοφυλακίς, etc., in the test data). Probably the remaining, less frequent forms could also be added automatically to Morpheus' lexicon, by detecting similar looking forms and assigning them to a paradigm based on their suffixes (e.g. because the genitive ποτομαφυλακίδος *potamofulakídos* and the nominative ποταμοφυλακίς *potamofulakís* both occur in the papyrus data, we can deduce that the stem is *potamofulakí-s/dos*). The mismatch between training and test data also leads to problems on other linguistic levels than the lexicon, however. For instance, past indicative verb forms on *–ον* can be either analyzed as first person singular or third person plural. In the tagging results, I found a couple of instances in which a first-person singular was incorrectly tagged as a third person plural, and no examples of the opposite. This is probably because the lexical probabilities of the tagger are calculated on the basis of literary Greek, in which first-person verbs are less frequent than, for instance, in papyrus letters. A possible solution is to give the in-domain data a larger weight than the out-of-domain data during training, or to bring in some information about the test data during the training process, e.g. by using word vector representations (see Section 2). Another possibility would be to tag the papyri completely unsupervisedly, although the amount of tokens (about 4.5 million) is likely to be too small for this (see Piotrowski, 2012: 89).

There are still many possibilities to improve syntactic parsing accuracy (e.g. testing other parsing models and approaches, improving the quality of the training data, finding strategies that can deal with structures that are difficult to parse such as coordination). As for morphological tagging, joint morphological and syntactic analysis is likely the most productive step to further increase accuracy, as I argued above. At any rate, while this step will likely have a significant effect, some remaining problems are still difficult to resolve. The choice

between first-person singular and third-person plural, for instance, often depends on complex semantic and pragmatic world knowledge regarding which actions are more likely to be performed by the speaker and which by other people in a given communicative situation (see also Manning, 2011). Although such issues should be in theory resolvable (as humans are, after all, able to do this), they may well be too complex to solve for the current generation of part-of-speech taggers.

Acknowledgements

The author would like to thank his PhD supervisors, Dirk Speelman, Toon Van Hal, and Mark Depauw, as well as two anonymous reviewers for their constructive feedback which has greatly helped to improve this article. The present study was supported by the Research Foundation Flanders (FWO) [grant number: 1162017N].

References

Acedański, S. (2010). A morphosyntactic brill tagger for inflectional languages. In *Advances in Natural Language Processing*. Springer: Reykjavik, pp. 3–14.

Adafre, S. F. (2005). Part of speech tagging for Amharic using conditional random fields. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, MI, June 2005.

Ballesteros, M. and Bohnet, B. (2014). Automatic feature selection for agenda-based dependency parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, August 2014.

Ballesteros, M. and Nivre, J. (2012). MaltOptimizer: an optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, April 2012.

Bamman, D. and Crane, G. (2011). The ancient Greek and Latin dependency treebanks. In Sporleder, C., van den Bosch, A., and Zervanou, K. (eds), *Language Technology for Cultural Heritage*. Heidelberg: Springer, pp. 79–98.

Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Sydney, July 2006.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, July 2006.

Bohnet, B. *et al.* (2013). Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, **1**: 415–428.

Cayless, H. A. *et al.* (2016). *Idp.data: Data from the Integrating Digital Papyrology Project.* papyri.info. http://github.com/papyri/idp.data.

Celano, G. G. A., Crane, G., and Majidi, S. (2016). Part of speech tagging for ancient greek. *Open Linguistics*, **2**(1): 393–9.

Cohen, S. B. and Smith, N. A. (2007). Joint morphological and syntactic disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, June 2007.

Covington, M. A. (1990). Parsing discontinuous constituents in dependency grammar. *Computational Linguistics*, **16**(4): 234–236.

Crane, G. (1991). Generating and parsing classical greek. *Literary and Linguistic Computing*, **6**(4): 243–245.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, Portland, OR, June 2011.

Daumé, H. III. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, June 2007.

Denis, P. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, Vol. 1, Hong Kong, December 2009.

Denniston, J. D. (1978). *The Greek Particles*. Oxford: Clarendon.

Dik, H. (2007). *Word Order in Greek Tragic Dialogue*. Oxford: Oxford University Press.

Dik, H. and Whaling, R. (2008). Bootstrapping classical Greek morphology. Paper presented at *Digital Humanities 2008*, Oulu, June 2008.

Ekbal, A., Haque, R., and Bandyopadhyay, S. (2008). Maximum entropy based Bengali part of speech tagging. *Advances in Natural Language Processing and Applications, Research in Computing Science*, **33**: 67–78.

Gimpel, K. *et al.* (2011). Part-of-speech tagging for twitter: annotation, features, and experiments. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, Vol. 2, Stroudsburg, PA, June 2011.

Goldwater, S. and Griffiths, T. (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, June 2007.

Habash, N. and Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, MI, June 2005.

Hajič, J. (2000). Morphological tagging: data vs. dictionaries. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, Seattle, WA, May 2000.

Hajič, J. and Hladká, B. (1998). Tagging inflective languages: prediction of morphological categories for a rich, structured tagset. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Vol. 1, Montreal, August 1998.

Haug, D. T. and Jøhndal, M. (2008). Creating a parallel treebank of the old Indo-European bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakech, June 2008.

Keersmaekers, A. and Depauw, M. (forthcoming). *Bringing Together Linguistics and Social History in Automated Text Analysis of Greek Papyri*. Classics@ 18, Heidelberg.

Kraft, R. (1988). *Morphologically Analyzed Septuagint. Computer-Assisted Tools for Septuagint Studies (CATSS)*. University of Pennsylvania. http://ccat.sas.upenn.edu/gopher/.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML, Williamstown, MA, July 2001.

Lee, J., Naradowsky, J., and Smith, D. A. (2011). A discriminative model for joint morphological disambiguation and dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, Portland, OR, June 2011.

Liddell, H. G. and Scott, R. (1940). *A Greek-English Lexicon. Revised and Augmented throughout by Sir Henry Stuart Jones with the Assistance of Roderick McKenzie*. Oxford: Clarendon Press.

Mambrini, F. and Passarotti, M. C. (2012). Will a parser overtake Achilles? First experiments on parsing the ancient Greek dependency treebank. In *Proceedings of the 11th Workshop on Treebanks and Linguistic Theories (TLT11)*, Lisbon, December 2012.

Manning, C. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, Tokyo, February 2011.

McDonald, R. *et al.* (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, October 2005.

McDonald, R. and Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, June 2007.

McDonald, R. and Satta, G. (2007). On the complexity of non-projective data-driven dependency parsing. In *Proceedings of the 10th International Conference on Parsing Technologies*, Prague, June 2007.

Müller, T. *et al.* (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, September 2015.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, October 2013.

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. In *Proceedings of the Joint*

Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol. 1, Suntec, August 2009.

Nivre, J. *et al.* (2016). Universal dependencies v1: a multiling[illegible] treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, May 2016.

Nivre, J. *et al.* (2007). MaltParser: a language-independent syst[illegible]or data-driven dependency parsing. *Natural Language Engineering*, **13**(2): 95–135.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. San Rafael, CA: Morgan & Claypool.

Porter, S. E. and O'Donnell, M. (2010). Building and examining linguistic phenomena in a corpus of representative papyri. In Evans, T. V. and Obbink, D. D. (eds), *The Language of the Papyri*. Oxford: Oxford University Press, pp. 287–311.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, May 1996.

Sawalha, M. and Atwell, E. S. (2010). Fine-grain morphological analyzer and part-of-speech tagger for Arabic text. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valleta, May 2010.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In International Conference on New Methods in Language Processing*, Manchester, 1994.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Vol. 1, Manchester, August 2008.

Schnabel, T. and Schütze, H. (2014). Flors: fast and simple domain adaptation for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, **2**: 15–26.

Stolk, J. V. (2016). Scribal and phraseological variation in legal formulas: the adjectival participle of *γπάρχω*+ dative or genitive pronoun. *Journal of Juristic Papyrology*, **45**: 255–290.

Tauber, J. K. (ed.) (2017). MorphGNT: SBLGNT Edition [Data Set]. https://github.com/morphgnt/sblgnt.

Tsarfaty, R. *et al.* (2010). Statistical parsing of morphologically r[illegible] languages (SPMRL): what, how and whither. In *Proceedings of the First Workshop on Statistical Parsing of Morphologically Rich Languages*, Los Angeles, CA, June 2010.

Turner, E. G. (1980). *Greek Papyri: An Introduction*. Oxford: Clarendon.

Vierros, M. and Henriksson, E. (2017). Preprocessing Greek Papyri for linguistic annotation. *Journal of Data Mining and Digital Humanities, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages*.

Notes

1 The term 'papyri' is used in the field of papyrology not only to refer to texts written on papyri but also to texts written on potsherds, wood, parchment, etc., i.e. 'all materials carrying writing in ink done by a pen' (Turner, 1980). I only include documentary, i.e. non-literary papyrus texts in my analysis.

2 A searchable interface for the project described in this article can be consulted at http://www.trismegistos.org/words.

3 http://www.trismegistos.org. Trismegistos does not only cover Greek papyri but also cover papyri in other languages such as Latin, Demotic, and Coptic, as well as epigraphical sources.

4 This information is also available in the HGV (*Heidelberger Gesamtverzeichnis*, http://aquila.zaw.uni-heidelberg.de) database, although especially the genre information is represented in a rather different way—on the basis of 'keywords', which do not only contain text type (e.g. letter, petition, and list) but also content information.

5 The Trismegistos unique identifier (abbreviated as TM) will be used in this article to refer to specific papyrus texts.

6 https://github.com/PerseusDL/morpheus.

7 In some cases the amount of morphological information that is expressed in a single word can become quite high: Ancient Greek participles, for instance, express number, gender, case, tense/aspect, and voice, so that there are more than 150 possible participle forms for a given verb.

8 Excluding parts-of-speech, my tag set includes thirty-three morphological features for Greek to be taken into account during part-of-speech tagging, lemmatization, and syntactic parsing.

9 The same problem arises with syntactic parsing, since not all morphological features may be equally beneficial to determine syntactic relationships as well (Tsarfaty *et al.*, 2010). Hence some widely used dependency parsers such as MaltParser (Ballesteros and Nivre, 2012) and Mate

(Ballesteros and Bohnet, 2014) have developed feature selection algorithms to ensure that the most informative features for syntactic parsing out of a given training set will be used.

10 On the other hand, obviously Greek word order is not completely random: some words can only occur in a certain place in the clause (e.g. subordinate conjunctions at the beginning of the clause), and the order of words within syntactic constituents is typically far more predictable (Dik, 2007). Therefore a radical NLP approach to Ancient Greek that completely ignores word order would likely perform quite poorly as well.

11 As for dependency parsing, languages with a flexible word order often contain a significant amount of non-projective arcs, i.e. structures with crossing dependency edges (or, in other words, discontinuous structures) (McDonald and Satta, 2007). In fact, the amount of non-projective arcs in Ancient Greek might be exceptionally high: the data from the Ancient Greek Treebanks (Mambrini and Passarotti, 2012) use contains 22–27% non-projective arcs. Since such non-projective structures are difficult to handle for a dependency parser, specialized algorithms are needed for these structures (Covington, 1990; McDonald *et al.*, 2005; Nivre, 2009).

12 I did not look into this topic at the moment due to time and space constraints.

13 It is, however, fair to say that there are some complications concerning orthographic conventions. Frequent combinations of particles are often written together, e.g. *méntoi* for *mén* + *toi*. I decided to regard these combinations as a single word, since the meaning can often not be derived compositionally (Denniston, 1978). Some function words such as articles and the conjunction *kaí* ('and') ending on a vowel often contract with the following word when this word starts with a vowel (a phenomenon known as crasis), e.g. *kamoí* for *kaí* + *emoí* ('to me'). For the time being, such contractions were given the tag of the word having the highest degree of semantic content—in this case *personal pronoun* + *singular* + *common gender* + *dative* (the tag of *emoí*)—although it might be preferable to divide these combinations into two tokens.

14 It is not always easy to separate spelling and morphological corrections, however. There are some clear-cut cases such as ἔχι ékhi for ἔχει ékhei (TM 77953)—where –i is not a possible Greek suffix, so the editor inevitably regularized the spelling—and Ψάϊς Psaïs for Ψάτος Psaïtos (TM 73518)—where the substitution of -s for –tos is not pholologically plausible, so the editor clearly corrected the nominative Psaïs to the genitive Psaïtos. In other cases, however, it is more dubious to say whether the scribe used a non-standard spelling or a non-standard grammatical form, e.g. the form υἱῶ huiõ (TM 28410), corrected by the editor to υἱοῦ huioũ, where the scribe could have used the dative suffix õ instead of the genitive suffix oũ either because of grammatical reasons or because of confusion in the spelling of the oũ-sound. In such cases, the best possible option is to provide both the 'original' tag 'dative' and the 'corrected' tag 'genitive', so that the corpus user can decide which version would be most suitable for their research.

A deficiency of this method is the fact that editorial regularizations are not always done consistently (since papyrus texts have been edited by several different editors). This was not particularly problematic for part-of-speech tagging (since non-regularized spellings were infrequent enough to not have a significant effect on tagging accuracy) but should be taken into account when using the corpus data.

15 Editors usually regularize linguistic forms for two reasons: (1) because the form used is inconsistent with the language use at the time, e.g. the use of the nominative case for the object in the transitive construction, or (2) to bring the language usage closer to an earlier (usually Classical) Greek norm, e.g. the verb ὑπάρχω was used with a dative complement in Classical Greek but could also be used with a genitive complement is Postclassical Greek, and as a consequence, the genitive in papyri is often corrected to a dative by the editor (Stolk, 2016). In either case, using the regularized form would be beneficial for syntactic parsing: in the case of (1), the correction would respond more closely to the syntactic structure (we would want to analyze a nominative used with object function as an 'object' as well), while in the case of (2), the papyrus data would correspond better to the (literary, often Classical Greek) training data.

16 This number only includes evaluated tokens, i.e. no punctuation marks or incomplete words due to physical damage to the papyrus, so the actual token count is a little higher. The following texts are included: TM 701, 739, 961, 1,732, 1,872, 3,342, 3,346, 5,364, 7,126, 8,810, 11,099, 11,453, 14,145, 18,048, 19,702, 20,620, 22,021, 23,875, 29,702, 30,617, 36,009, 36,090, 36,197, 36,707, 37,205, 88,690, 129,772, 140,178, and 144,995.

17 I was able to increase RFTagger's accuracy with 0.5% (to 95.1%) by using a 6-gram instead of a 3-gram model. Since the other taggers take much more time to train, no additional parameters were tested. For all

other tests described in this section, I used a 3-gram model.

18 More precisely, some word classes were different—I assigned participles and infinitives to unique word classes instead of considering them as a verbal 'mood', and divided pronouns into several subclasses instead of considering them as adjectives—and I also made some minor changes within morphological categories (e.g. 'medio-passive' present and perfect verbs were called 'middle').

19 I was forced to make only use of data that was both morphologically and syntactically annotated, i.e. the prose data encompassed in the AGDT and PROIEL projects. This implies that the Septuagint was excluded, and the New Testament of the PROIEL instead of the MorphGNT project was used.

20 Mate for instance made twenty-one mistakes involving the confusion between nominative and accusative, while RFTagger made 22 and MarMoT 25.

21 The possible values for these categories are the following:

Part-of-speech: Noun, adjective, verb, article, personal pronoun, demonstrative pronoun, indefinite pronoun, relative pronoun, interrogative pronoun, numeral, adverb, preposition, conjunction, particle, and interjection.

Derivative category: Infinitive and participle.

Number: Singular and plural.

Voice: Active, middle, and passive.

Tense: Present, aorist, imperfect, future, perfect, and pluperfect.

Degree (only adjectives): Positive, comparative, and superlative.

Case: Nominative, vocative, accusative, genitive, and dative.

Person: 1, 2, and 3.

Mood: Indicative, subjunctive, optative, and imperative.

Gender: Masculine, feminine, and neuter.

The possible values 'dual' for number and 'future perfect' for tense also exist, but these are rare in the training corpus and non-existent in the test corpus.

22 I do not have an explanation why Plato and especially Aesop have a negative impact on tagging accuracy; however, since the effect is relatively small (a 0.05% and 0.14% drop in accuracy, respectively) and they both contribute to less than 1% of the training data, this could simply be a coincidence.

23 It is not surprising that these structures are analyzed so badly, since coordination structures, which code symmetric relationships, are obviously hard to represent with asymmetric dependencies. This is a general problem for dependency parsing in any language rather than for Greek alone, although the free word order of Ancient Greek certainly complicates the matter further. In this respect, a possible way to improve parsing accuracy would be changing the way coordination structures are annotated in the training data, since McDonald and Nivre (2007) show that parsing precision for these structures can range considerably (from 40% to 80%) depending on the way in which they are represented. Some early experiments show that for Ancient Greek as well LAS for nodes in coordination structures in particular can improve with about 24% (from 48% to 72%) using some of the annotations McDonald and Nivre (2007) propose.

24 However, since I tagged documentary papyrus texts, which have a more rigid word order than literary Greek texts, it is not clear how significant this problem would be when tagging literary Greek.