Mechanical Part of Speech Tagging for the Latin Language

Digital Humanities: Practice & Theory

Fergus J. P. Walsh[*]

Monday 3rd May, Easter Term, 2021

1 Introduction

> *Grammatica est scientia recte loquendi, et origo et fundamentum liberalium litterarum.*
>
> Isidore of Seville *Etymologiae*

From Antiquity until the present day, much of Latin grammar instruction has remained the same: the correct identification (parsing) or production of accidence, that is the inflectional forms of Latin words, the meanings of these inflections and thus a word's role in a sentence.

In addition to being foundational to a student's own development in Latin proficiecy, parsing is also foundational to the commentary tradition within Classical studies. Today both print commentaries and digital reading environments provide morphological aid to Latin students.[1] However, not every Classical text is served by a commentary, while in the domain of post-Antique Latin literature the *desideratum* of subsiduary materials is even greater.[2] Today also, new textbooks and students' materials that require vocabulaires and notes are being produced at a greater rate than ever before.[3]

Despite this, context-specific part of speech tagging is still in its infancy. In the past year, two neural network taggers have been relased, but this enquiry will use the hidden Markov model RFTagger to tag Latin texts, while also treating the available

[*] The author wishes to thank John Morris Esq. for technical assistance with Linux and the RFTagger, and Dr. Mair Lloyd (Cambridge Schools Classics Project) for providing access to the tagged Eduqas GCSE texts. This paper is also accompanied by some of the data sets employed in this enquiry and a Jupyter Notebook detailing part of the data analysis. Word count: 3,077.

[1] Gibson 2021, pp. 179, 185, 196; Crane and Bamman 2007, p. 35.

[2] Knight and Tilg 2015, pp. 3–4; "the quantity of neo-Latin literary material is enormous." Moul 2017, pp. 1, 7.

[3] Piazza 2017 and the Latin Novella Database: `latinnovelladatabase.blogspot.com`

manually-tagged Latin corpora suitable for use in such a machine learning application.[4]

2 Background

2.1 Latin morphology

> In speche be these .viii. partes folovvyung: nowne, pronowne, verbe, participle; declyned. Aduerbe, coniunction, preposition, interecition; vndeclyned.
>
> William Lily *An Introdvction of the Eighte Partes of Speche* (1542)

Nearly every Latin word has multiple inflectional forms which designate its grammatical role in a sentence. A noun and pronoun will belong to one of 36 possible combinations of number, case and grammatical gender; an adjective one of 108; a finite verb will be one of 432;[5] while participles will be one of a staggering 648 possible combinations of number, tense, voice, gender and case.

However, some inflections are shared across different forms of a word. Hence the token *veni* could either be the first person singular perfect active indicative or the second person present imperative active of the lemma *venio* (“I have come” or “come!”).[6] Moreover, morphological forms can even be shared across different parts of speech. Consider the token *malo*, which depending on a given context could be either a verb, two forms of two adjectives or two forms of three nouns:

malo	1st person singular present active indicative of	*malo,* “I prefer”
	masculine dative singular	*malus, -a, -um,* “to/for an evil man”
	masculine ablative singular	*malus, -a, -um,* “by/with/from an evil man”
	neuter dative singular	*malus, -a, -um,* “to/for an evil thing”
	neuter ablative singular	*malus, -a, -um,* “by/with/from an evil thing”
	masculine dative singular	*malus,* “to/for a mast”
	masculine ablative singular	*malus,* “by/with/from a mast”
	neuter dative singular	*malum,* “to/for an apple”
	neuter ablative singular	*malum,* “by/with/from an apple”
	feminine dative singular	*malus,* “to/for an appletree”
	feminine ablative singular	*malus,* “by/with/from an appletree”

[4] Lee Min-cheol 2020; Clérice 2020.

[5] Distinguishing between passive and deponent verbs.

[6] Discussion of the significance of vowel length is delayed until section 6.

2.2 Other Latin language taggers

Human readers use context and meaning to disambiguate homonyms, yet in the case of machine parsing, merely identifying the correct form or forms of a token is not enough to disambiguate.[7] In their testing of the LatMor tagger, Springmann, Schmid and Najock report that on average a given token has 2.5 possible analyses and that, "disambiguation of the set of possible morphological analyses to the single analysis that is correct within a given sentence context is not a task for morphology."[8]

One way to deal with the ambiguity of homonyms is to default to always choosing the most frequent lemma. The Collatinus program, which includes a lemmatisation and tagging tool, orders amiguous lemmata by frequency in the LASLA corpus (see section 3).[9] In their stylistic analysis of Statius *Achilleid*, Heslin likewise defaulted to lemmatising a token by its most frequent lemma, assuming that a longer entry in Lewis and Short corresponds to a higher word frequency.[10]

Although both of these cases are attempts at disambiguing for the purpose of lemmatisation, rather than part of speech tagging, they are unsatisfactory. A contextual tagger is needed, but both Heslin and Ouvrard and Verkerk doubt that a probabilistic tagger using a hidden Markov model is an appropriate approach for part of speech tagging in Latin.[11] They wrongly assume that a hidden Markov model tagger cannot be effective when used in conjuction with a large tagset (most tagsets are 50–150 tags large, whereas the LASLA corpus has a tagset of over 3,000) or in the case of a language with a variable word order.[12]

However, following the breakout success of the Trigrams'n'Tags (TnT) tagger, two taggers have been released that modify TnT for morphologically rich languages: HunPos for Hungarian and the RFTagger for German and Czech (with out-of-the-box support for Russian, Slovak and Slovene).[13] Like Latin, these languages have a variable word order and a complex inflectional patterns.[14] Schmid and Laws tested the RFTagger on the German Tiger Treebank, which has a tagset of 700, and reported accuracies of up to 92%. On the Czech Academic Corpus, with a tagset of 1,200, the RFTagger could acheive an accuracies of 90%.[15]

[7] "Where a form is ambiguous, Morpheus simply spits out multiple alternative parses. This is entirely appropriate when the output is designed to be read by a human." Heslin 2019, p. 392.

[8] Springmann, Schmid and Najock 2016, p. 389.

[9] Ouvrard and Verkerk 2009–2016; Ouvrard and Verkerk 2019, pp. 7–8; Ouvrard and Verkerk 2014.

[10] ”This is a quick and dirty tactic." Heslin 2019, p. 396.

[11] "The obvious problem with applying such a model to Latin is that it is predicated on the assumption that word order is strongly correlated with meaning." Heslin 2019, p. 394; "Very high accuracies are obtained with modern languages, where the order of the words in the sentence is rather fixed. It is not demonstrated that the same fidelity can be reached with Latin, where the order of the words is free, or at least much freer than in modern languages." Ouvrard and Verkerk 2019, p. 19.

[12] "The method relies of the hypothesis that the sequences of tags are characteristic of the language." Ouvrard and Verkerk 2019, p. 18; "it is not obvious that relying upon strict word-order to provide sematic cues is a sound general assumption for literary Latin of the classical period." Heslin 2019, p. 394.

[13] Brants 2000; "Even without a formal survey it is clear that TnT is used widely in research labs throughout the world." Halácsy, Kornai and Oravecz 2007, p. 1; Schmid and Laws 2008a; Schmid and Laws 2008b.

[14] Halácsy, Kornai and Oravecz 2007, p. 209; Schmid and Laws 2008a, p. 1.

[15] Schmid and Laws 2008a, pp. 5–7.

Treebank	Approximate Time Span	Tokens
LASLA	3rd cent. B.C. – 1st cent. A.D.	1,630,825
PROIEL Latin Treebank	1st cent. B.C. – 5th cent. A.D.	200,163
Latin Dependency Treebank 2.1	1st cent. B.C. – 4th cent. A.D.	29,138
Late Latin Charter Treebank 2	774–897 A.D.	242,411
Index Thomisticus Treebank 2.7	1225–1274 A.D.	450,515

Table 1: Latin language tagged treebanks.

2.3 The RFTagger

The RFTagger can clearly overcome the supposed limitations of the hidden Markov model. For a given sequence of n words $w_1 \dots w_n$, the tagger calculates the most probable sequence of tags $t_1 \dots t_n$ as:

$$\hat{t}_{1:n} = \underset{t_1 \dots t_n}{\arg\max}\, P(t_1 \dots t_n | w_1 \dots w_n) \tag{1}$$

The tagger makes two assumptions: firstly that the probability of a word appearing is independent of other words and tags, and secondly that the probability of a tag is dependent only on the c previous tags (where c is a tuneable parameter: $c = 2$ is a trigram). Thus the probability that tag sequence $t_1 \dots t_n$ is paired with $w_1 \dots w_n$ is:[16]

$$\prod_{i=1}^{n} P(t_i | t_{i-k}^{i-1}) P(w_i | t_i) \tag{2}$$

The RFTagger, however, decomposes the tags into simple attributes and uses decision trees to estimate the probability of each attribute in the tag independently.[17] Thus for a Latin noun, three decision trees are in operation: one for number, one for case and one for gender, each calculating the probability of a tag as a product of the other attribute probabilities. In this way, the tagger can handle the large tagsets required, where backoff smoothing methods and discriminateively trained taggers fail.[18]

3 Latin Treebanks

A machine learning approach necessitates a sufficently large corpus of manually tagged texts with which to train the RFTagger, and so it is necessary here to briefly outline the Latin language treebanks currently available (Table 1).

Mention has already been made of the LASLA treebank, a corpus of approximately 1,500,000 tokens.[19] Unfortunately the LASLA corpus is not open access and has not been used in this enquiry.

[16] For further details, see: Jurafsky and Martin 2020, pp. 157–158; Schmid and Laws 2008a, p. 1.

[17] Schmid and Laws 2008a, p. 3.

[18] Schmid and Laws 2008a, p. 2.

[19] See `web.philo.ulg.ac.be/lasla/textes-latins-traites/` and Denooz 2004.

Part of Speech	Person	Number	Tense	Mood	Voice	Gender	Case	Degree
`n` noun	`1` first	`s` singular	`p` present	`i` indicative	`a` active	`m` masculine	`n` nominative	`p` positive
`v` verb	`2` second	`p` plural	`i` imperfect	`s` subjunctive	`p` passive	`f` feminine	`g` genitive	`c` comparative
`a` adjective	`3` third		`r` perfect	`n` infinitive	`d` deponent	`n` neuter	`d` dative	`s` superlative
`d` adverb			`l` pluperfect	`m` imperative			`a` accusative	
`c` conjunction			`t` future perfect	`p` participle			`v` vocative	
`r` adposition			`f` future	`d` gerund			`b` ablative	
`p` pronoun				`g` gerundive			`l` locative	
`m` numeral								
`i` interjection								
`e` exclaimation								
`u` punctuation								

Table 2: Perseus Digital Library Latin Dependency Treebank tagset.

The Latin Dependency Treebank, part of the Perseus Project, is a treebank that contains Classical prose and verse literature and part of the Book of Revelation from St. Jerome's Latin Vulgate.[20] The tagset consists of nine characters, encoding part of speech, person, number, tense, mood, voice, gender, case and degree respectively (Table 2).[21] Moreover, every token is represented by a nine character-long tag, even if not all features are applicable (a `-` is used in such an instance). For example, the first line of Propertius *Elegiae* I is tagged as:

(1) *Cynthia* *prima* *suis* *miserum* *me* *cepit*
`n-s---fn- m-s---fn- p-p---mb- a-s---ma- p-s---ma- v3sria---`
ocellis ,
`n-p---mb- c--------`

This tagset was designed for Latin and Ancient Greek, and uses a taxonomy that has remained unchanged since the time of Servius, Donatus or Priscian. However, the other three treebanks are part of the wider Universal Dependencies project, which shepherds and publishes annotated treebanks from a wide range of languages.[22] Universal Dependencies treebanks use a revised Stanford dependency representation, which aim to be able to describe the grammar of any natural language.[23] Thus new features absent from the Classical grammatical tradition are added, terms to describe features unique to Latin or Ancient Greek are absent, while different instances of a single lemma may be split across multiple word classes, where they would not be in the Perseus treebank.

The PROIEL (Pragmatic Resources in Old Indo-European Languages) project has published a Latin treebank that includes St. Jerome's New Testament, selections from Cicero and Caesar, in addition to the later works Palladius *Opus Agriculturae* (5th century A.D.) and *Peregrinatio Aetheriae* (4th century A.D.)[24]

The *Index Thomisticus* Treebank consists of St. Thomas Aquinas' *Summa Contra Gentiles.*[25] Winge did not include the *Index Thomisticus* when training the taggers for

[20] Currently available at `github.com/PerseusDL/treebank_data` where the Readme also includes a partial description of texts covered.

[21] Crane and Bamman 2007, p. 38.

[22] See `universaldependencies.org/introduction.html`

[23] de Marneffe et al. 2014.

[24] See `proiel.github.io` and Dag and Jøhndal 2008. At the time of writing, the works of Plautus and Terence are also currently being added to the treebank.

[25] See `github.com/UniversalDependencies/UD_Latin-ITTB`, Cecchini et al. 2018 and Passarotti

the Macronizer because, “it is restricted to a single author, and only covers mediaeval Latin.”[26] Although I later discovered through personal correspondence that the exclusion of the *Index Thomisticus* was due to the problem of converting the tagset (it did not use the Universal Dependency tagset at that time) rather than any animus against St. Thomas or Medieval Latin in general, it is worth noting here that Winge’s attitude is not unusual among Classicists. Sadly the requirement for brevity prevents me from proving this assertion with any scientific rigour, but Latin literature of the Medieval period does not represent a “dialect” different from Classical Latin. At the level of morphology and the nominal cases, the usage of Medieval writers, particularly those as erudite as St. Thomas, did not differ in any meaningful way from that of the writers of Antiquity.[27] There is no reason, therefore, not to use the *Index Thomisticus* in this enquiry.

The Late Latin Charter Treebank is a body of legal documents, written by both professional and unprofessional scribes in Tuscany during the 8th and 9th centuries A.D.[28] These texts are an especially important resource in Latin and Romance historical linguistics; they are written in Latin, not in the emerging Tuscan vernacular, but show clear departures from both the Classical and later Medieval norms of lexis, orthography, morphology and syntax.[29] In particular, nominal declension patterns sometimes differ due to phonological change or a shift in a noun’s gender.[30]

As mentioned above, as part of the Universal Dependencies project, the tagsets of the PROIEL, Late Latin Charter Treebank and *Index Thomisticus* treebanks are quite consistent. Below is an example from the *Index Thomisiticus* (development fold sentence 8 *“finis enim est bonum uniuscuiusque.”*):

```
(2) finis                                           enim
    NOUN Case=Nom|Degree=Pos|Gender=Masc|Number=Sing ADV advmod
    est
    AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
    bonum
    NOUN Case=Nom|Degree=Pos|Gender=Neut|Number=Sing
    uniuscuiusque
    PRON Case=Gen|Degree=Pos|Gender=Neut|Number=Sing|PronType=Ind
    .
    PUNCT punct
```

and Dell’Orletta 2010

[26] Winge 2015, p. 10.

[27] “there is nothing in the inflectional system of Medieval Latin…that would have disconcerted a Roman writer of the classical period.” “The degree of classicism would depend merely on the extent to which a writer was familar with, and eager to imitate, classical style.” Rigg 1996, pp. 83, 89.

[28] Originally published by D. Barsocchini in *Memorie e documenti per servire all’istoria del Ducato di Lucca* (1818–1841). Korkiakangas 2021, pp. 1–2, 6–7

[29] “it is postulated that the writers did not conceptually separate the language they wrote from the spoken language of the time…the Latin of LLCT reflects the uninterrupted evolution of the spoken Latin.” Korkiakangas 2021, p. 2; Cecchini, Korkiakangas and Passarotti 2020, p. 933; “The continuum between the spoken and written languages seems to have lasted longer in Italy than in France.” Clackson and Horrocks 2007, pp. 267–269.

[30] See: Korkiakangas and Passarotti 2011, pp. 110–113; Korkiakangas 2021, pp. 4–5; and Clackson and Horrocks 2007, p. 278.

These four open access treebanks amount to 922,227 tokens in total, comparable with available treebank corpora for some modern languages.[31]

4 Method

> Anybody can write a Latin analyzer over a weekend...
>
> Gregory Crane in Springmann 2015

Four models were created: one trained on the combined PROIEL and Perseus treebanks, one on the Late Latin Charter Treebank, one on the *Index Thomisiticus* and finally one all four treebanks combined. For each model, the context parameter *c* was set between 2 and 10 (i.e. unigram to 9-gram) and so nine varaint parameter files for each model were generated. These parameter files are labelled CL02–CL10, LL02–10, ML02–10 and Max02-10 respectively.

For test data, the *Index Thomisiticus* test fold was kept back (31,748 tokens, approximately 7% of the whole corpus) and two short extracts from Classical literature were used: the myth of Ceyx and Alcyone from Ovid *Metamorphoses* 11 and Tacitus' narration of Boudica's revolt (*Annales* 14). These two texts are the prescribed texts for the 2022–23 Eduquas GCSE Latin exam, and were manually tagged by Dr. Mair Lloyd for the online Explorer reading environment.[32] Although only 700 tokens long each, it was necessary to source additional texts due to the relative paucity of Classical literature available in the treebanks. A test fold from the Late Latin Charter Treebank was not used due to the requirements of brevity.

The major part of the data pre-preparation was converting the three Universal Dependencies tag sets and the proprietary CSCP tag set into the Perseus tagset. This was chosen due to its compact nature and its consistent length of 9 characters accross all tokens. As can be seen from the companion notebook, the tags for each token were extracted from the CoNLL-U file and a Perseus tag generated by searching the original tag with regular expressions. The Universal Dependencies tag set does not taxonomise morphology in the same way as the Perseus tag set. Verbs such as *sum* ("I am"), *habeo* ("I have") and *fio* ("I become") are classed either as verbs or auxiliaries dependent on context. Words such as *qui, quae, quod* ("who"), *hic, haec, hoc* ("this") and *iste, ista, istud* ("that of yours"), which Classicists describe as pronouns, are termed determiners.[33]

Frustratingly, the PROIEL treebank contained 465 tokens that were not parsed or lemmatised due to being Greek or Hebrew words or calendar expressions (merely given the tag "X"), and as a quick fix they were all classed as nouns. If time had

[31] In fact, the Universal Dependencies project only has an English corpus of 648,000 tokens! `universaldependencies.org/treebanks/en-comparison.html`

[32] See `www.exams.cambridgescp.com/Array/eduqas-component-3a-latin-literature-narratives` for the prescribed texts and Explorer reading environment. I wish to thank the Cambridge School Classics Project again for use of these files. Due to copyright restrictions, I am unable to share the manually tagged data set as part of the submission.

[33] Gildersleeve and Lodge 1867 class *qui*, *quidam*, *hic*, *iste*, *ipse* as pronouns, as does the Perseus treebank.

allowed, I would have manually tagged these tokens myself, as Hebrew and Greek terms do regularly occur in Latin texts, particularly in Christian theological works, which account for a very significant portion of the Latin patrimony.[34]

The Perseus tagset has a tripartite classification of verbal voice: active, passive and deponent. A deponent verb is active in meaning but exhibits passive morphology and are common in Latin texts.[35] However, the Universal Dependencies tagset classes all deponent verbs as passive.[36] Fortunately, deponent verbs can easily be identified by their lemma, which all treebanks paired with each token and tag. Using a list of deponent verbs from Wiktionary, verbs tagged as passive but with their lemma in the Wiktionary list were reclassified as deponent.[37]

Finaly, the Perseus treebank contained a number of typos, the GCSE texts required verbal mood to be added and both the GCSE and PROIEL texts required the addition of punctuation. The RFTagger requires a line break after each sentence, and so these were added after each full stop, question mark and exclaimation mark. The RFTagger also requires a full stop between each tag in the tagset (`n-s---fn-` becomes `n.-.s.-.-.-.f.n.-`)

To overcome the problem of sparsity in the training data, both HunPos and the RFTagger allow the use of an external lexicon in the training process. As part of the Macronizer, Winge prepared an external lexicon derived from the output of Morpheus: every inflected form from Lewis and Short 1879 with a corresponding tag.[38] This lexicon contains the alternative spellings with "J" for "I" and "U" for "V", which removed the need to regularise the orthography in the treebanks.

5 Results

The accuracy of the models' predictions are shown in Table 3 and Figure 1. These were calculated by comparing the predicted tag to the ground truth using a simple equals true/false test; thus a single mistake in one of the nine attributes leads to a predicted tag being categorised as incorrect.

It is immediately apparent that the test scores vary significantly by model and by test set, but that the context parameter has virtually no impact on a models' predictive performance. Secondly, that the worst scores result from the two very small GCSE test sets, whereas on the much larger *Index Thomisiticus* test set, the scores are much higher. The model trained on all four treebanks (Max), has the most consistent performance, whereas the model trained on the Perseus and PROIEL corpora (CL) performs just as well on the two Classical test sets, but poorly on the *Summa Contra Gentiles.* Here ML, the model trained on the other fold of that text, excels, but does performs middingly on the Classical test sets. In all cases LL, the model trained on

[34] See `stats.xml` file in `github.com/proiel/proiel-treebank/releases/tag/20180408`

[35] Gildersleeve and Lodge 1867, pp. 85–87; "Overall 7% of the verbs in a text are deponents…Some of these verbs are very common." Pinkster 2015, p. 283.

[36] Conversely, the Classical Language Toolkit classes them all as active.

[37] `en.wiktionary.org/wiki/Category:Latin_deponent_verbs` The list appears to be based on the headwords of Lewis and Short, but this has not been verified.

[38] The `rftagger-lexicon.txt` file at `github.com/Alatius/latin-macronizer`. The nature of the file was confirmed via personal correspondence.

Model	Boudica	Ceyx	*I.T.*	Model	Boudica	Ceyx	*I.T.*
CL02	66.6	74.7	70.1	ML02	63.5	70.9	89.0
CL03	67.2	75.5	70.3	ML03	63.3	71.6	89.0
CL04	67.2	75.7	70.3	ML04	64.2	71.3	89.0
CL05	67.3	75.0	70.3	ML05	64.6	72.2	89.0
CL06	67.3	75.2	70.2	ML06	64.1	71.8	89.0
CL07	67.0	75.1	70.3	ML07	63.8	72.2	89.1
CL08	67.2	75.0	70.3	ML08	63.3	72.0	89.0
CL09	66.8	75.7	70.2	ML09	63.3	72.0	89.0
CL10	67.2	75.4	70.2	ML10	63.5	72.2	89.0
LL02	56.5	65.5	66.6	Max02	65.9	75.0	87.2
LL03	57.9	65.8	66.5	Max03	66.5	76.1	87.4
LL04	57.6	64.0	66.3	Max04	65.8	76.2	87.4
LL05	56.9	64.3	66.2	Max05	65.0	75.9	87.4
LL06	59.6	63.6	66.3	Max06	65.8	76.4	87.4
LL07	59.1	63.8	66.3	Max07	66.5	76.2	87.3
LL08	59.2	64.0	66.3	Max08	66.3	75.7	87.3
LL09	59.1	63.6	66.2	Max09	66.5	76.1	87.2
LL10	58.9	63.6	66.2	Max10	66.2	76.4	87.2

Table 3: Whole-tag accuracies for the four models with variable context parameter. Figures given as percentages to 1 d.p.

the Late Latin Treebank does poorly, though on the Boudica test set an improvement in performance can be observed with the context parameter at 5-grams or greater. It is also noteworthy that all models performed better on the Ceyx test set; though not marked in the data itself, this text was originally a hexameter poem, while the *Annales*, the source text of the Boudica data set, is a prose text.

6 Discussion

These results, while not comprehensive or conclusive, are promising: firstly, the RFTagger can achieve a classification performance on Latin texts comparable with its performance on other modern languages, showing that a hidden Markov model tagger can be sucessfully applied to Latin language parsing. Secondly, it has been shown that the type of Latin text used to train the tagger does have an effect on the tagger's performance; while the Max model's consistent performance shows that both a large and diverse training set leads to a more generalised tagger.

It was also notable that the context parameter had no effect on the tagger's performance at all. I had hypothesised that a larger context window would be yield better results.[39] It was not possible to optimise the tagger's other parameters here, but the

[39] On the Czech Academic Corpus, the context parameter likewise had very little effect on the RFTagger's performance, though on the German test set results improved with a larger *n*-gram value Schmid and Laws 2008a, p. 6.

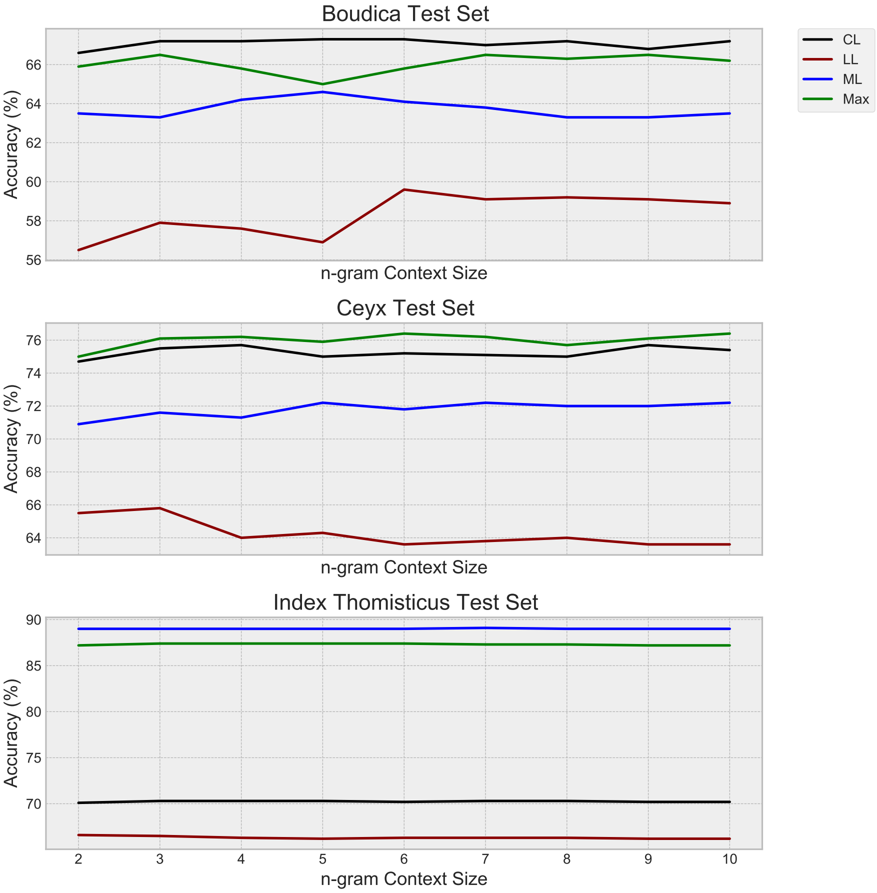

Figure 1: Whole-tag accuracies for the four models with variable context parameter.

pruning criterion and drop-out rate of the decision trees can also be changed by the user.

Throughout this whole enquiry there has been an elephant in the room: vowel length. In Latin each of the seven vowels occurs either as a short or a long vowel. Often these vowels are phonemic; the contrasting vowel lengths encode meaning and are used to distniguish apparent homonyms. Earlier, it was stated that *veni* could either be "I have come" or "come!"

$$veni \begin{cases} v\bar{e}n\bar{\imath} & \text{1st person perfect active indicative, "I have come"} \\ ven\bar{\imath} & \text{second person present active imperative, "come!"} \end{cases}$$

However, the perfect stem of the verb has a long *ē*, while the present stem has a short *e*. In speech, the two are not homophones, while if vowel length were marked in print, then the ambiguity would be absent. Unfortunately, both print editions and digitised texts of Latin works do not print vowel length. Not only is this a great shame for all Latinists, but in the field of natural language processing, needless ambiguity has been generated. Moreover, this issue is hardly discussed in the literature, despite its extreme pertinence.[40]

Clearly there is more work to be done in this area, not least in comparing the performance of different taggers or employing back-off methods to harness the performance of non-context specific taggers. It is exciting that Latin language treebanks now are so numerous and cover a broad range of texts, though the Medieval and Early Modern periods remain underrepresented. There has not been the opportunity in this enquiry to discuss the applications of mechanical part of speech tagging beyond producing grammatical commentaries, vocabularies or the Macronizer, but I hope that future Latin studies will be able to profit from the advances in treebanking and context-specific tagging.

[40] LatMor is perhaps the only tagger to take vowel length into account, but if the texts it encounters do not have macra, then that advantage goes to waste. Springmann, Schmid and Najock 2016, p. 387.

References

Brants, T. (2000). "TnT: A Statistical Part-of-Speech Tagger". In: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. URL: arxiv.org/pdf/cs/0003055.pdf.

Cecchini, F., Korkiakangas, T. and Passarotti, M. (2020). "A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages". In: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC2020)*.

Cecchini, F., Passarotti, M., Marongiu, P. and Zemanl, D. (2018). "Challenges in Converting the *Index Thomisticus* treebank into Universal Dependencies". In: *Proceedings of the Universal Dependencies Workshop 2018 (UDW 2018)*.

Clackson, J. and Horrocks, G. (2007). *The Blackwell History of the Latin Language*. Malden, M.A.: Wiley.

Clérice, T. (Sept. 2020). *Deucalion Latin Lemmatizer*. Version 0.0.3. DOI: 10.5281/zenodo.4043059. URL: https://doi.org/10.5281/zenodo.4043059.

Crane, G. and Bamman, D. (2007). "The Latin Dependency Treebank in a Cultural Heritage Digital Library". In: *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pp. 33–40.

Dag, 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)*. Ed. by C. Sporleder and K. Ribarov, pp. 27–34.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. and Manning, C. (2014). "Universal Stanford Dependencies: A cross-linguistic typology". In: *Proceedings of LREC*. URL: nlp.stanford.edu/pubs/USD_LREC14_paper_camera_ready.pdf.

Denooz, J. (2004). "Opera latina : une base de données sur internet". In: *Euprhosyne* 32, pp. 79–88. URL: web.philo.ulg.ac.be/lasla/wp-content/uploads/sites/7/2019/02/WEBLasla.pdf.

Gibson, R. (2021). "Fifty Years of Green and Yellow: The Cambridge Greek and Latin Classics Series 1970–2020". In: *Classical Scholarship and Its History: From the Renaissance to the Present. Essays in Honour of Christopher Stray*. Ed. by S. Harrison and C. Pelling. Berlin: De Gruyter, pp. 175–218. DOI: doi.org/10.1515/9783110719215.

Gildersleeve, B. and Lodge, G. (1867). *Latin Grammar*. 1997 Bristol Classical Press reprint. London: Macmillan.

Halácsy, P., Kornai, A. and Oravecz, C. (2007). "HunPos: an open source trigram tagger". In: *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 209–212.

Heslin, P. (2019). "Lemmatizing Latin and Quantifying the 'Achilleid'". In: *Intertextuality in Flavian Epic Poetry*. Ed. by N. Coffee, C. Forstall, L. Galli Milic and D. Nelis. Berlin: De Gruyter, pp. 389–408.

Jurafsky, D. and Martin, J. (2020). *Speech and Language Processing*. 3rd ed. Draft of 30 December 2020. URL: web.stanford.edu/~jurafsky/slp3.

Knight, S. and Tilg, S. (2015). "Introduction". In: *The Oxford Handbook of Neo-Latin*. Ed. by S. Knight and S. Tilg. Oxford: Oxford University Press, pp. 1–9.

Korkiakangas, T. (2021). "Late Latin Charter Treebank: contents and annotation". In: *Corpora* 16. In press. URL: researchportal.helsinki.fi/en/publications/late-latin-charter-treebank-contents-and-annotation.

Korkiakangas, T. and Passarotti, M. (2011). "Challenges in Annotating Medieval Latin Charters". In: *Journal for Language Technology and Computational Linguistics* 26.2, pp. 105–116.

Lee Min-cheol (15th Oct. 2020). *Lamon, The Latin PoS Tagger & Lemmatizer.* Version 0.2.0. URL: github.com/bab2min/lamonpy.

Lewis, C. and Short, C., eds. (1879). *A Latin Dictionary.* New York: Harper.

Moul, V. (2017). "Introduction". In: *A Guide to Neo-Latin Literature.* Ed. by V. Moul. Cambridge, U.K.: Cambridge University Press, pp. 1–14.

Ouvrard, Y. and Verkerk, P. (2009–2016). *Collatinus: Linguae Latinae Lemmatizatio.* Version 11.2. URL: outils.biblissima.fr/en/collatinus.

— (2014). "Collatinus, un outil polymorphe pour l'étude du latin". In: *Archivum Latinitatis Medii Aevi* 72, pp. 305–311.

— (2019). "Collatinus & Eulexis : Latin & Greek Dictionaries in the Digital Ages." In press. URL: hal.archives-ouvertes.fr/hal-02385036.

Passarotti, M. and Dell'Orletta, F. (2010). "Improvements in parsing the Index Thomisticus treebank. Revision, combination and a feature model for medieval Latin". In: *Training* 2, pp. 61–024.

Piazza, J. (2017). "Beginner Latin Novels, a General Overview". In: *Teaching Classical Languages* 8.2, pp. 154–166. URL: tcl.camws.org/sites/default/files/TCL8.2Piazza.pdf.

Pinkster, H. (2015). *The Oxford Latin Syntax.* Vol. 1. Oxford: Oxford University Press.

Rigg, A. (1996). "Morphology and Syntax". In: *Medieval Latin: An Introduction and Bibliographic Guide.* Ed. by F. Mantello and A. Rigg. Washington, D.C.: Catholic University of America Press, pp. 83–92.

Schmid, H. and Laws, F. (2008a). "Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained PoS Tagging". In: *COLING.*

— (2008b). *RFTagger.* URL: www.cis.lmu.de/~schmid/tools/RFTagger.

Springmann, U. (2015). *Treebanking ancient languages — current and prospective research* (slides). Universität Leipzig. URL: springmann.net/talks/2015-12-15-Leipzig-latmor.pdf.

Springmann, U., Schmid, H. and Najock, D. (2016). "LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity". In: *Open Linguistics* 2.1, pp. 386–392. DOI: doi.org/10.1515/opli-2016-0019. URL: www.cis.lmu.de/~schmid/tools/LatMor.

Winge, J. (June 2015). "Automatic Annotation of Latin Vowel Length". Bachelor's Thesis in Language Technology. Uppsala, Sweden: Department of Linguistics & Philology, Uppsala University. URL: http://stp.lingfil.uu.se/exarb/arch/winge2015.pdf.