Arne Skjærholt

More, faster: Accelerated corpus annotation with statistical taggers

We present our experiments with annotating a Latin corpus using an assisted annotation procedure where the corpus to be annotated is pre-annotated by a statistical tagger. This assisted procedure gives a notable reduction in annotator error compared to the unassisted annotation of previous annotation efforts, even with a huge tagset (1 000 tags) and modest tagger accuracy due to limited training data and domain effects.

1 Introduction

When creating corpora of richly inflecting languages like Latin, the most time-consuming (and boring) task is morphological annotation. Qualified labour for a classical language is also hard to come by, adding unneeded strain to already limited budgets and slowing down the pace of corpus development. Thus, we would very much like to speed up this process.

We present here a simple, effective and cheap way of achieving this. Using almost no custom components, relying instead on off-the-shelf software, we leverage an existing Latin corpus to accelerate annotation of a new text with the help of an HMM tagger whose output is corrected by annotators. Even though the tagger is far from the 95% accuracy of the state-of-the-art in tagging in general and the tagset is extremely large (more than 20 times larger than the Penn Treebank tagset), we improve both the speed and error rate of manual annotation considerably.

After a quick review of related research, both into statistical taggers for Latin and annotation assisted by taggers, we present the corpus used for our experiments and its annotation procedure, as well as a brief outline of the particularities of Latin as a language. Then we present the details of our experiments: the accuracy of the tagger itself and how it compares to previous taggers for Latin, the effect of assisted annotation on annotation speed, and its effect on annotator error. Finally, we sum up our conclusions from these experiments and mention some possible avenues for future research.

1.1 Previous work

The automatic analysis of Latin morphology has been the subject of a few previous studies. Poudat and Longrée (2009) used the LASLA corpus[1] to explore the automatic analysis of Latin morphology with HMMs, and the influence of factors such as author,

[1] http://www.cipl.ulg.ac.be/Lasla/

genre and time period on tagging performance. Skjærholt (2011) used the PROIEL corpus (more on this in the next section) to compare the viability of HMMs versus the more sophisticated CRF models and studied the possibility of using constrained decoding to increase tagger performance on out-of-domain data. Bamman and Crane (2008) and Passarotti (2010) also explore statistical tagging of Latin in the greater context of developing lexical resources; they both achieve comparable results to the more in-depth studies, Bamman and Crane (2008) using TreeTagger (Schmid, 1994) and Passarotti (2010) using HunPos (Halácsy et al., 2007).

The basic idea of using a computer to generate output to be post-processed is essentially the same as Bar-Hillel's (1960) suggestion that machine translation output be corrected by a human translator and the translator's amanuensis proposed in Kay (1997). The concrete idea of letting human annotators correct tagger output rather than starting from scratch is a common one, and has been explored in depth by several studies. When creating the Penn Treebank, Marcus et al. (1993) found that manual tagging took twice as long and gave double the inter-annotator disagreement compared to correcting tagger output, a position largely corroborated by Fort and Sagot's (2010) more in-depth study of the influence of tagger accuracy. However, their experiments used the Penn Treebank, and we need to verify that their results still hold with a tagset as large as ours. Furthermore, Fort and Sagot used in-domain data to train their tagger, whereas we use out-of-domain data, which means that we require more data to get similar tagger performance. Dandapat et al. (2009) are guardedly optimistic, but they too show that correcting tagger output yields better data than annotating from scratch.

2 Language & corpus

2.1 A crash course in Latin

Latin is an Indo-European language with a long and interesting history, ancestor of the Romance languages of today. The very first traces of Latin language date back to eight century BCE, and the oldest literature to survive until our day, the comedies of Plautus, date to 200 BCE or thereabouts. The language is typical of classical Indo-European languages of roughly the same age, such as Ancient Greek and Sanskrit; it is a richly inflecting language with synthetic morphology and a large array of forms and morphemes, even though the Latin system represents a radical departure and restructuring of the ancestral system better preserved in Greek and Sanskrit.

The history of Latin is usually divided into several periods, the most important being Classical Latin, which dates from around the 1st century BCE to the first century CE; most of the Latin authors commonly known today, like Caesar, Cicero, Virgil and Horace, belong to this era. Anything earlier than Classical Latin is counted as Old Latin. After the Classical era, the language starts to split into two languages: Vulgar Latin, the language of the people, already quite different from the literary language in the Classical era, becomes more and more a separate language, eventually becoming the Romance languages. The literary language on the other hand remains relatively unchanged for several centuries.

Roughly speaking, the morphology of Latin (which is the interesting part, vis-a-vis the present work) can be divided into two largely independent parts: nominal inflection and verbal inflection. The verbal system governs all finite forms of the verb, while the nominal system covers inflection of the remaining infinite forms of the verb, nouns, and adjectives. A few words fall outside of these two groups, most notably the pronouns. Both the nominal and verbal systems are further subdivided into five declinations and four conjugations, respectively; these subdivisions are again more or less independent, forming the different forms with different morphemes, especially in the case of the nominal system. Finally, it is quite common for several forms of a word to be identical, especially in the nominal system.

2.2 Corpus

For our experiments, we used the Pragmatic Resources of Old Indo-European Languages[2] (PROIEL) corpus to train the tagger whose output the annotators correct and to gather data from unassisted annotation to compare our assisted approach with. The PROIEL project aims to study the pragmatics of several classical IE languages (Ancient Greek, Old Church Slavic, Classical Armenian, Gothic, and Latin) by creating a large parallel corpus of several such languages, to allow for large-scale contrastive analysis. The main part of the corpus is the translation[3] of the New Testament in the respective language, but some other texts from the various languages are included as well.

The corpus is morphologically annotated with two tagsets: a part-of-speech (PoS) tagset for features belonging to the lemma, and a morpho-syntactic descriptor (MSD) tagset for features that vary according to the form of the word. The PoS tagset is relatively coarse with only 23 tags, corresponding to the ten parts of speech of traditional grammar, augmented with finer subdivisions for some parts of speech (most notably nine kinds of pronoun) and a foreign word class; the full list is given in table 1. The MSD tagset is a fixed-width format ten characters wide, where each position corresponds to a particular morphological feature such as case, mood or tense; a list of tags relevant to Latin is presented in table 2. The PoS tag is attached to the lemma of a word, and MSD tags to each token in the corpus, and in total there are 962 distinct PoS-MSD pairs in the Latin part of the corpus. The syntactic annotation is in the style of dependency grammar, with the addition of secondary dependencies to fill the external roles of open functions, similar to structure sharing in LFG and HPSG (Haug, 2010, 1).

The PROIEL annotation procedure is somewhat idiosyncratic; instead of each sentence being annotated by two independent annotators and then resolving any disagreements, the PROIEL annotation procedure is in two steps: First an annotator (graduate students, for the most part) analyses the morphology and syntax of the sentence. The

[2] `http://foni.uio.no:3000/`

[3] Or original, in the case of ancient Greek

Tag	Meaning	Tag	Meaning
A-	adjective	Pc	reciprocal pronoun
C-	conjunction	Pd	demonstrative pronoun
Df	adverb	Pi	interrogative pronoun
Dq	relative adverb	Pk	personal reflexive pronoun
Du	interrogative adverb	Pp	personal pronoun
F-	foreign word	Pr	relative pronoun
G-	subjunction	Ps	possessive pronoun
I-	interjection	Pt	possessive reflexive pronoun
Ma	cardinal numeral	Px	indefinite pronoun
Mo	ordinal numeral	R-	preposition
Nb	common noun	V-	verb
Ne	proper noun		

Table 1: PoS tagset

Position	Feature	Values
1	Person	1st (1), 2nd (2), 3rd (3)
2	Number	singular (s), plural (p)
3	Tense	present (p), imperfect (i), future (f), perfect (r), pluperfect (l), future perfect (t)
4	Mood	indicative (i), subjunctive (s), imperative (m), infinitive (n), participle (p), gerund (d), gerundive (g), supine (s)
5	Voice	active (a), passive (p)
6	Gender	masculine (m), feminine (f), neuter (n), m/n (o), m/f (p), m/f/n (q), f/n (r)
7	Case	nominative (n), vocative (v), accusative (a), genitive (g), dative (d), ablative (b)
8	Degree	positive (p), comparative (c), superlative (s)
9	Unused[a]	—
10	Inflection	inflecting (i), non-inflecting (n)

[a] Used for strong/weak inflection in Gothic and Old Church Slavonic

Table 2: MSD tagset

Corpus	Sentences	Tokens	Avg. tok/sen
BG	1 417	26 663	18.8
Vulgata	12 459	112 135	9.0
Peregrinatio	921	17 553	19.1

Table 3: Annotated corpus sizes at the time of writing

sentence is then reviewed by a more senior annotator to ensure that the analysis is correct (Haug and Jøhndal, 2008, 27–28).

The Latin part of the corpus is comprised of three texts: the *Vulgata* translation of the Bible, Caesar's *Bellum Gallicum* (*BG*) and *Peregrinatio Aetheriae*, a 5th century Vulgar Latin account of a pilgrimage to the Holy Land. Of the three, the *Vulgata* corpus is by far the largest at more than 100 000 annotated tokens, with the *BG* corpus at 25 000 tokens. Detailed statistics are given in table 3. We omit the *Peregrinato* corpus from the experiments in the present work, since the Vulgar Latin of this text is simply too different from the Classical Latin of Caesar and the literary style of Jerome's *Vulgata*. In particular, the restructuring of the rich morphological system of Latin into the more modest Romance system is well under way, which means that many inflections are used in ways that are flat out wrong in the more classically informed Latin.

Many students' first encounter with Latin is Caesar's *BG*, and its opening sentences will serve us well as an example:

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit.

In all, Gaul is divided in three. Of these, the Belgians inhabit one, the Aquitans another, and those who are called Celts in their own language, or Gauls in our own, inhabit the third. All of them differ between each other in language, traditions and laws. The river Garonne separates Gauls from Aquitans, and the Seine and Marne from the Belgians.

The morphological annotation of the opening of the first Latin sentence, corresponding to the first sentence of the translation, is shown in figure 1; the first two characters correspond to the PoS tags of table 1 and the remaining ten to the MSD tagset of table 2. Thus, *in* is a preposition (`R–`) which is indeclinable (`–––––––––n`) and *divisa* the feminine nominative singular of the perfect passive participle (`–srppfn––i`) of a verb (`V–`).

Finally, a brief word on how PROIEL compares to the LASLA corpus. First of all, the latter project started in 1961, which means that the size of the corpus is quite

Gallia	est	omnis	divisa
Ne-s---fn--i	V-3spia----i	Px-s---pn--i	V--srppfn--i
in	partes	tres	[quarum . . .]
R----------n	Nb-p---fa--i	Ma-p---pa--i	

Figure 1: Morphological annotation of *BG* 1.1.1

significant: 1.6 million words[4], an impressive figure compared to the 140 000 words in the PROIEL corpus. The LASLA corpus is primarily morphologically annotated however, only a limited amount of information related to verbs is annotated. Second, the LASLA tagset is quite a bit larger than the PROIEL tagset, due to its encoding of inflectional classes in the PoS part of the tagset; Poudat and Longrée (2009) report a total of 3 732 distinct tags, more than three times the 960 tags in the PROIEL corpus. Unfortunately, the raw data of the LASLA corpus aren't publicly available, so we couldn't use it to train our tagger.

Index Thomisticus and the Perseus Latin Dependency Treebank are two other publicly available corpora of Latin. In general, more training data for a statistical model is a good thing, but we decided not to use these sources. Index Thomisticus is a treebank of the works of Thomas Aquinas, and is medieval Latin and excluded on the same basis as *Peregrinatio*. The Perseus LDT is a 50 000 token treebank, made up of selections of various important Latin author's work. Linguistically, these texts are a good fit with our training corpus, but unfortunately the Perseus and PROIEL tagsets are not the same, and converting from the Perseus tagset to PROIEL is a non-trivial task which would most likely introduce quite a bit of noise to our data.

3 Assisted annotation

In order to investigate the properties of assisted annotation and its efficacy for the annotation of Latin, we selected a new text for annotation. The text to be annotated is Cicero's *Epistulae ad Atticum* (*Att*), a collection of letters to his friend Titus Pomponius Atticus. This is a fairly large corpus, composed of 4 561 sentences and 61 193 tokens (giving an average of 13.4 tokens per sentence). Linguistically and stylistically, this text is most closely aligned with Caesar's *BG* rather than the later (and simpler) *Vulgata* text.

At the outset, the primary objective of this new assisted annotation procedure is to provide a faster annotation rate compared to the unassisted annotation. It would also be nice if the assisted annotation results in better annotation (that is, fewer errors) compared to the unassisted procedure. We will quantify both of these dimensions. Inter-annotator agreement is another standard measure of annotation quality; we will

[4] `http://www.cipl.ulg.ac.be/Lasla/tlatins.html`, retrieved 5/9/2011

not quantify this, for the simple reason that it is not possible to do so with our present dataset, since each sentence is only annotated by a single annotator.

3.1 The tagger

The tagger we used for our experiments is Thorsten Brants' Trigrams'n'Tags (TnT), described in Brants (2000). TnT is a fairly straightforward trigram HMM tagger, but with one important addition: the unknown word model. Instead of estimating emission probabilities of words not seen in training by some kind of discounting strategy, the suffixes of words seen in training are matched against the suffixes of the unknown word, and the emission probability of the longest matching suffix is used as the word's emission probability. This strategy works very well for Latin, since it's exclusively suffix-inflecting.

The model used to pre-process the *Att.* corpus was trained on the concatenation of the *Vulgata* and *BG* corpora, using TnT's default options. The tagger output was then combined with a partial finite-state morphology that was made available to us, such that if TnT's analysis was one of those licenced by the morphology, the lemma of the finite-state analysis was added as well. If the analyses did not match, the lemma was set to "FIXME".

A brief interview with the annotators who have annotated the new corpus makes it clear that the preprocessed corpus at least makes the annotation work more bearable. However, we would like to quantify the effects of preprocessing the corpus with the statistical taggers as well. An important first datum is simply the raw performance of the tagger, and how this compares to previous results. Both Poudat and Longrée (2009) and Skjærholt (2011) evaluate tagger performance on in-domain and out-of-domain (OOD) data, and despite the important differences between the two corpora, obtain remarkably similar results, summarised in table 4. In particular, their results result in an overall accuracy of 84.3%, even though Poudat and Longrée's training corpus is larger than Skjærholt's. Given this, it seems reasonable to believe that our results using the PROIEL data can be meaningfully compared to Poudat and Longrée's other results. The accuracy of 76.9% is encouraging and somewhat unexpected, given previous tagging results; before evaluating the tagger we expected a result closer to the 63% of row *e*. The *c* result of 77.2% in table 4 is very close to our final accuracy, but the training corpus in that experiment was quite a bit larger: 352 820 tokens compared to our 139 620.

Instead of using a statistical tagger, another option would be to use a rule-based tagger, such as Words by William Whitaker[5], Morpheus (Crane, 1991), originally developed for classical Greek and later adapted to Latin, or Lemlat (Passarotti, 2000). However, such a tagger outputs all possible analyses for an ambiguous token, which doesn't fit very well with the DB schema of the annotation tool, and it is preferred that the annotators correct a single analysis rather than choose from a potentially long list of options.

[5] http://ablemedia.com/ctcweb/showcase/whitakerwords.html

Experiment	TA	OOV	IV
Poudat and Longrée (2009)[a]	84.3	?	?
Poudat and Longrée (2009)[b]	63.7	?	?
Poudat and Longrée (2009)[c]	77.2	?	?
Skjærholt (2011)[d]	84.3	60.7	88.9
Skjærholt (2011)[e]	62.8	33.3	85.0
Vulgata & *BG* on *Att*	76.9	50.0	85.7

[a] LASLA, *BG* books 1–2,4–7 on book 3

[b] LASLA, *BG* and *Bellum Civile* on *1st Catilinarian*

[c] LASLA, historical texts on *1st Catilinarian*

[d] PROIEL, *BG* 10-fold cross-validation

[e] PROIEL, trained on *BG*, tested on *Vulgata*

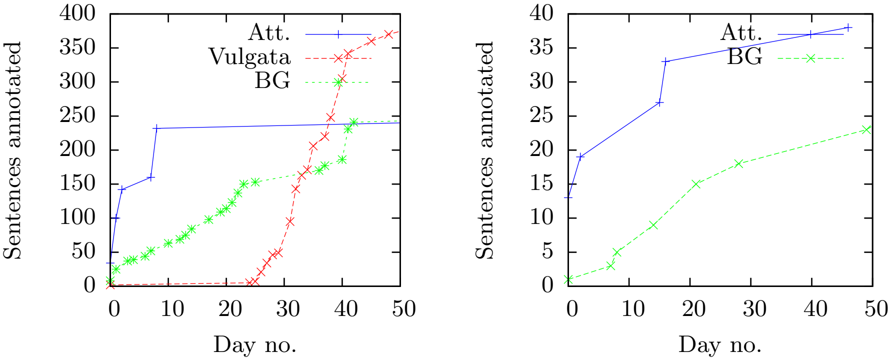
Table 4: Tagging accuracy (in percent) on Latin. Token accuracy (TA), out-of-vocabulary (OOV) and in-vocabulary (IV) accuracy.

3.2 Annotation speed

Annotation speed is harder to gauge accurately. The only information available from the PROIEL DB dump is a timestamp, the date and time when the annotator saved the annotation. Thus, information like time spent per sentence or how many sentences were annotated in a single sitting is not explicitly represented in the data. One could conceivably synthesise a number of sentences per sitting, for example by grouping sentences annotated within some threshold, say five or ten minutes, as belonging to the same annotation session. We have not done this however. Instead we truncate the timestamps to the date portion and count only the number of sentences annotated by each annotator each day. No matter which approach is used to synthesise such a statistic from the timestamps would involve drawing some arbitrary line in the sand, and we believe this approach to be the least problematic approach.

There are two annotators, Aulus and Gaius, both master's students, who have annotated the *Att.* corpus. Aulus has done the majority of the annotation (420 sentences) while Gaius has annotated 38 sentences. Aulus is the more experienced annotator, having previously annotated 4 207 sentences of the *Vulgata* corpus and 852 sentences in *BG*, while Gaius has annotated 332 sentences of *BG* before the *Att.* annotation. This gives us something to compare the *Att.* effort with. Figures 2 (a) and (b) show the total number of sentences by each annotator along the y-axis and number of days since his first annotation along the x-axis.

In the case of Aulus, it seems quite clear that annotation of the *Att.* corpus is helped by the pre-annotation; compared to the *BG* corpus, 10 days of annotation has produced what took 40 days without pre-annotation. The *Vulgata* graph has roughly the same slope, without pre-annotation, as the *Att.* effort, but this text is significantly simpler



(a) Aulus

(b) Gaius

Figure 2: Annotation speeds. Note that the y-axes of the two figures are different, due to the difference in number of sentences annotated.

and not as comparable to Cicero's text as Caesar's. For Gaius the picture is less clear, but it seems that the assisted annotation of the *Att.* corpus is a bit faster than the unassisted annotation of *BG*.

A more rigorous way to test these notions is to apply hypothesis testing to the data. Table 5 shows the sufficient statistics to apply Student's t-test to the dataset: the number of datapoints (n), the mean number of sentences annotated per day (μ) and the sample standard deviation (s). Applying the t-test to these data, we find that Aulus' annotation of *Att.* is significantly different from his annotation of *BG*, but not his annotation of *Vulgata*, nor are Gaius' two annotation series significantly different (all at the 95% level).

3.3 Annotator error

Another important metric is annotator error. To quantify this we extract all the sentences that have been reviewed by the expert annotator (as per the PROIEL annotation procedure outlined in section 2.2); of Aulus' sentences 25 have been reviewed, and 32 of Gaius'. Using audit data tracking changes done to the tokens, we counted the number of tokens whose morphology had been changed since the sentence was annotated, the results of which are summarised in table 6. The numbers presented are token error (TE), sentence error (SE), the number of sentences with errors (n), the mean number of mistagged tokens per sentence with errors (μ) and the sample standard deviation of the number of tokens per mistagged sentence (s).

In the case of Aulus, we cannot meaningfully compare the number of errors per mistagged sentence in *Att.* with the others since there are only two such sentences, each

Annotator	n	μ	s
Aulus, *Vulgata*	96	43.8	39.8
Aulus, *BG*	55	15.5	13.6
Aulus, *Att*	9	46.7	23.6
Gaius, *BG*	44	7.55	4.61
Gaius, *Att*	5	7.60	3.21

Table 5: Annotation statistics. Number of days with annotation (n), number of sentences annotated per day mean (μ) and standard deviation.

Annotator	TE	SE	n	μ	s
Aulus, *Vulgata*	2.80	18.8	545	1.28	0.628
Aulus, *BG*	8.27	70.3	415	2.17	1.35
Aulus, *Att*	0.529	8.00	2	1.00	0.00
Gaius, *BG*	7.44	66.9	222	2.52	1.89
Gaius, *Att*	2.11	9.38	3	2.67	1.53

Table 6: Annotator error. Token error (TE) and sentence error (SE) in percent, number of mistagged sentences (n), number of erroneous tokens per mistagged sentence mean (μ) and sample standard deviation (s).

with a single mistagged token; this gives a standard deviation of zero, which again means that no matter the confidence level, the bound on the mean will always be ±0. Gaius' data one the other hand, do not have this problem; his numbers of wrong tokens per mistagged sentence are not significantly different at the 95% level.

Taking a somewhat broader perspective yields a quite pleasing view as well. Even though the number of errors per sentence once an error is made appears to be relatively unchanged, the number of errors made is reduced dramatically: Annotation error at the token level is reduced by almost a factor of four, from 7.44% to 2.11%, for the junior annotator and more than a full order of magnitude to a mere half percent for the more experienced Aulus. Sentence-level error is likewise encouraging, reduced by almost an order of magnitude for both annotators, from the neighbourhood of 70% to slightly less than 10%.

4 Conclusion & future work

All in all, the results of our study are very encouraging. Our experienced annotator, Aulus, benefits the most from assisted annotation. His annotation speed increased dramatically, with our proxy for annotation speed tripling compared to the unassisted annotation of the *BG* corpus, which is comparable in terms of linguistic complexity; his error rate was reduced by an order of magnitude, on both token and sentence level. Gaius, the less experienced annotator, had no measurable change in annotation speed, but he too made far fewer errors both in terms of tokens and sentences.

Based on this evidence we believe that assisted annotation is an excellent tool, even for annotation tasks with huge tagsets, and that if data is available to train a tagger, the assisted approach is preferable to unassisted annotation, both in terms of annotator error and annotation speed. For annotation error both annotators had a sizeable decrease in error, but for speed only one of the two annotators showed an improvement; however, given that our proxy measure for speed was tripled in the case of Aulus, and the assisted value is almost two and a half standard deviations from the unassisted value, we believe this to be indicative of a real improvement.

4.1 Future work

This work is a good start, but questions remain that we would like to see answered. First of all, a more in-depth study of assisted annotation using this kind of large tagset would be welcome. We have obtained good preliminary data, but certain metrics are unavailable to us given the nature of our dataset; chief among these are inter-annotator agreement and direct measurement of annotation speed. It would also be interesting to investigate further the influence of tagger accuracy on the various metrics. Fort and Sagot (2010) suggest that tagger accuracy in the 66–82% range is sufficient, and our tagger accuracy in the high seventies is consistent with these results; but since their work uses the fairly small Penn Treebank tagset one should verify that their results hold for extremely large tagsets as well. We would also like to investigate further if the kinds of error the annotators make differ qualitatively from the errors made with unassisted annotation.

References

Bamman, D. and Crane, G. (2008). Building a Dynamic Lexicon from a Digital Library. In Larsen, R., editor, *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries1*, pages 11–20, New York. ACM.

Bar-Hillel, Y. (1960). The Present Status of Automatic Translation of Languages. *Advances in Computers*, 1:91–163.

Brants, T. (2000). TnT: a statistical part-of-speech tagger. In Nirenburg, S., editor, *Proceedings of the sixth conference on applied natural language processing3*, ANLC '00, pages 224–231, Stroudsburg. Association for Computational Linguistics.

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

Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009). Complex Linguistic Annotation - No Easy Way Out ! A Case from Bangla and Hindi POS Labeling Tasks. In Stede, M. and Huang, C.-R., editors, *Proceedings of the third linguistic annotation workshop*, pages 10–18, Stroudsburg. Association for Computational Linguistics.

Fort, K. and Sagot, B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. In Xue, N. and Poesio, M., editors, *Proceedings of the fourth linguistic annotation workshop*, pages 56–63, Stroudsburg. Association for Computational Linguistics.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 209–212, Stroudsburg. Association for Computational Linguistics.

Haug, D. T. T. (2010). PROIEL Guidelines for Annotation.

Haug, D. T. T. and Jøhndal, M. L. (2008). Creating a parallel treebank of the old Indo-Eurpoean Bible translations. In Sporleder, C. and Ribarov, K., editors, *Proceedings of the Sixth International Language Resources and Evaluation1*, pages 27–34.

Kay, M. (1997). The Proper Place of Men and Machines in Language Translation. *Machine Translation*, 12(1-2):3–23.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Passarotti, M. (2000). Development and perspectives of the Latin morphological analyser LEMLAT. *Linguistica Computazionale*, 3:397–414.

Passarotti, M. (2010). 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages Workshop programme Additional Referees. In Sarasola, K., Tyers, F. M., and Forcada, M. L., editors, *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*.

Poudat, C. and Longrée, D. (2009). Variations langagières et annotation morphosyntaxique du latin classique. *Traitement Automatique des Langues*, 50(2):129–148.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Skjærholt, A. (2011). *Ars flectandi: Automated morphological analysis of Latin*. Master's thesis, University of Oslo.