Yela Schauwecker, Achim Stein

Automatic Morphosyntactic and Dependency Annotation of the Anglo-Norman Text Database

Abstract Non-standardized languages are an immense challenge for automatic annotation. This paper discusses the case of Anglo-Norman (AN), which is the variety of Old French (OF) spoken and written in medieval England for over 300 years, until well after 1400. In addition to presenting the irregularities in, for example spelling, inflection and word-order that are also characteristic of OF, AN developed particular spelling variants, shows even less consistent case-marking and considerable diachronic variation between the earliest (c1112) and the latest (c1440) texts in the Anglo-Norman text database (Rothwell and Trotter 2005; henceforth "ANdb").

We present the first attempt to provide an automatic grammatical analysis of the ANdb. We applied machine-learning techniques combined with lexicon-driven tools that were trained on OF resources. This paper is organized according to the individual steps in the annotation process: section 1 gives a succinct overview of the historical context and some relevant linguistic peculiarities of AN. Section 2 deals with the automated graphical "normalisation" of the texts. We generated regularized spellings that temporarily substituted the graphical forms during the annotation process to improve the accuracy of lemmatisation, part-of-speech tagging, and dependency parsing. Section 3 describes how a dependency parser developed for Old French was applied to the normalised version of the AN data, and discusses the usefulness of the parsed output for historical syntactic research.

Keywords Dependency parsing, part of speech tagging, automatic spelling normalisation, Anglo-Norman, Old French historical corpora

1 Anglo-Norman

1.1 Timeline of French in England

When William the Conqueror arrived at Pevensey in 1066, he brought with him the variety of Old French (OF) that was spoken in Normandy. At the beginning, Norman OF was the dominant code in England, which influenced the less prestigious Middle English. But, a few generations later, French speakers were almost always mother-tongue speakers of English, so that Insular French was maintained by largely fluent bilinguals (Ingham 2012). In contrast to earlier assumptions, Ingham found evidence that Anglo-Norman (AN)[1] showed no signs of decline until the fifteenth century (see also Hunt 2004). Since evidence for the systematic teaching of French emerges only just before that point, the acquisition of French by anglophone speakers until then must have taken place via natural interaction with French speakers.

1.2 Some features of Insular French

Knowing the syntactical features of AN, and in particular those that set AN apart from (continental) OF, is crucial to understanding the additional difficulties automatic annotation has to cope with in the case of insular texts. However, their detailed description is beyond the scope, and the topic, of this paper. Therefore, we just give some examples for the sake of illustration.

Being originally a variety of OF, AN shares most of the characteristic features of this language. Among these, the absence of a standardized spelling, inconsistent word-order, the licensing of null-subjects (see Marchello-Nizia 2009, among others), all represent major difficulties in automatic linguistic annotation. However, our tools are trained on OF resources (see section 3), and therefore, it is on OF that they achieve best results.

When it comes to Anglo-Norman, the situation gets more difficult. Even bare numerical comparison can reveal the high level of syntactic complexity in AN compared to OF: texts in the *Syntactic Reference Corpus of Medieval French (SRCMF)*[2] contain 24,171 "sentences" within 266,870 tokens, thus equalling an average of 11.04 words per sentence. Compared to that, texts in the ANdb contain

1 A number of researchers prefer the term "Anglo-French". We agree, but because of the more technical scope of this contribution, and in order to avoid confusion, we will use "Anglo-Norman" throughout this text.

2 Calculations based on the version 0.91, March 8, 2016.

3,111,982 in 148,353 "sentences"[3], which equals an average of 21 words per "sentence". In addition to that, AN was spoken and written for about 400 years, and therefore shows much diachronic variation in itself between the first (c1112) and the latest texts in the ANdb (c1440).

Like OF, AN showed considerable graphical irregularity from the start. But in the case of AN, these irregularities increased considerably, as AN phonology underwent some profound changes by the later thirteenth century. Phonological contrasts that had been kept up in earlier times ceased to be respected by later generations of speakers (Ingham 2012: 160). This is, of course, at least partly reflected in the orthography. In addition, AN exhibits a number of atypical traits by which it is set apart from continental OF (Ingham 2010), many of which are highly relevant to syntactic annotation. For example, the contrast between strong and weak forms of pronouns ceases to be respected in many cases (Grant 1978: 36–7; Johnston 1961:xix; Ingham 2010), and direct and indirect object case-marking is confused in later texts (Grant 1978: 36, Johnston 1961: xix; Ingham 2010). To summarise, as these examples illustrate, AN diverges from OF in syntax as well as in phonology and orthography. As a consequence, there is a clear difference between the texts our tools were trained on and the AN sources they are applied to. The following sections illustrate the approaches we adopted in order to bridge this gap and the results we achieved.

1.3 Pre-processing of the Anglo-Norman text database (ANdb)

The Anglo-Norman text database was compiled in order to support the *Anglo-Norman Dictionary* project (AND, Rothwell and Trotter 2005). It is freely accessible on the internet via the *Anglo-Norman On-Line hub* (ANHub[4]). It contains 78 texts, from c1112 to c1440.[5] At this point it must be noted that providing a fully annotated version of the ANdb is clearly beyond our possibilities, as is often the case with low-resourced but richly documented languages. However, in the case of AN, additional difficulties have to be dealt with. As we said above, the enormous syntactic complexity of especially the later AN documents – a considerable amount of "sentences" contains 200 and more tokens – would make full annotation extremely time-consuming and error-prone. In addition to that,

3 As to the notion of „sentence" in the SRCMF cf. infra, section 1.3.

4 http://www.anglo-norman.net.

5 The data used for the annotation presented here were kindly provided by Geert de Wilde within a research collaboration between the *Anglo-Norman Dictionary* project (AND) and the project *Borrowing of Argument Structure in Contact Situations* (BASICS), funded by the Deutsche Forschungsgemeinschaft 2015–2018.

texts often contain English, French and Latin words all within the same sentence. Thus, annotating them represents a major challenge even for well-trained human annotators, let alone elaborating a verified "gold-standard" version of the corpus. Moreover, as to the texts themselves, it has to be taken into account that the ANdb is heterogeneous in many respects: it contains prose as well as verse-texts, dating from very different periods and reflecting very different states of the language. They deal with an immense variety of topics and represent different types of texts, such as legal documents and charters, court proceedings, works of religious edification, pedagogical texts, medicine books, works on plants and on astronomy, etc. In total, the data being as they are, it is hard to imagine what a reliable sample in order to elaborate a partial "gold standard" could possibly look like. For the same reasons, building specialized tools, e.g. by creating an AN tagger lexicon, was clearly not feasible.

Instead, we had to work with existing resources and tools. But what started out as the second best option eventually turned out to be a very effective low-cost approach to our data, especially because the performance increased additionally after applying a layer of normalisation to our data prior to tagging. And since we normalised to a contemporary Medieval language, i.e., OF, our tagset did not need to be adapted, thereby allowing straight-forward comparisons across both languages. This work is meant to be of mutual benefit to the AND project (and eventual follow-up projects) and to the BASICS project on medieval language contact likewise. This contribution presents a snapshot of the work in progress, and we will refer to this stage as version 0.2 of the annotated corpus. In what follows, we describe the steps leading up to this version.

The first step consisted in ignoring the non-French passages[6] in the corpus. We did so for two reasons, firstly because they would have hampered the function of our analysis tools, which were trained for Old French. And secondly, because non-French passages and editorial notes are of no particular interest for the BASICS project. The XML markup of the texts could be used to identify most of the non-French passages, but some of them remained in the data and could not be dealt with manually at this point.

The second step was the segmentation, i.e. word form tokenisation and sentence splitting. On both the lexical and the syntactic levels this task is not trivial, but it does have a strong influence on the accuracy of automatic annotation. Since we use machine-learning tools for both tagging and parsing, the best results are achieved if word tokenisation and sentence splitting matches as closely as possible the texts the tools were trained on. The part-of-speech tagger (*TreeTagger*, Schmid 1994) uses parameters containing a lexicon of graphical

6 Non-French passages were, for example, Latin sentences in the psalters, and English and Latin paragraphs in macharonic texts.

forms most of which are associated with a lemma, so matching the input forms with the lexicon is important not only for the prediction of part-of-speech tags, but also for successful lemmatisation. Some of these tokenisation issues will be explained in more detail in the following section.

The accuracy of syntactic parsing depends quite heavily on the correct prediction of part-of-speech tags (more than on lemmatisation: in fact, lemmatisation had not significantly improved parser accuracy in previous tests with Old French; see Stein 2014), so word form tokenisation is also relevant for parsing. Moreover, since the main task of the parser is to predict the structure of a "sentence" (or at least of syntactic units defined as the relevant segments for parsing), the units of the input (the ANdb) should ideally follow the sentence definition of the training corpus (the SRCMF, Stein & Prévost 2013). However, this would have meant manually applying the SRCMF guidelines for sentence segmentation to the ANdb, which was not feasible at this stage of the project. In SRCMF, the unit "sentence" is defined minimally, as a structure containing no more than one main verb (which entails for example that coordinated main clauses are separated). Previous tests had shown that a dependency parser encounters fewer problems when input units are too long than when they are too short. Since verse texts contain many lines that are only parts of sentences, often lacking a verb, we decided not to use lines as an input unit, but to apply the same principles as for prose texts, i.e., we defined the sentence boundaries based on the punctuation marks inserted by the editors of the texts. Compared to the SRCMF principles, this often results in units that are larger than a SRCMF "sentence", e.g. enumerations containing main verbs or coordinations of main clauses (which were separated in SRCMF, according to the guidelines on http://srcmf.org). Since the parser, trained on SRCMF, has never seen coordinated predicates on the level of the main clause, it reacts by predicting for one of the coordinated structures a seemingly arbitrary category, for example "SjPer" (personal Subject) as in (1):

(1) [Confessïon desfait [**SjPer** et runt [Obj Trestots les liens
confession undoes and cuts all the bands
[ModA ke pecchez fount]]]]
which sins make
'Confession undoes and cuts all the relations that sins create. (all1237cors78)

However, the internal structures of both sentences are parsed correctly, which means that for syntactic queries that target structures other than coordination the analysis is acceptable. Thus, defining larger sentence units is the preferred choice, since it avoids the risk of producing units too small for the parser to analyse.

After the pre-processing the original files of the ANdb (including the "normalisation" described in section 2, our corpus (as of version 0.2) contained 3,111,982 tokenised graphical forms in 148,353 "sentences". After the segmentation procedure punctuation marks were deleted (again because the OF tools were trained on texts without punctuation marks). They are not included in the count of tokenised forms.

2 Normalisation of Anglo-Norman

2.1 Why normalisation matters

Due to spelling anomalies and to certain decisions on behalf of the scientific editors of the texts we are dealing with, queries cannot reveal all relevant hits. For example, querying the ANdb in its first, non-"normalised" version for *enportent* 'they carry away' yields three hits, among others:

(2) Dampnedeu les maudit S'il enportent un dener.
God them curses if-they carry-away one dime.
'God will curse them if they carry away one (single) dime'.
(alexander, 4/4 12th ct., v607)

But there is (at least) one more, which is:

(3) preignent lour blee et lenportent
take.3.PL their wheat and it_carry-away.3.PL
'They take their wheat and carry it away' (1419, Liber Albus, 783)

This fourth occurrence is not found because the editor chose to not intervene on agglutinated articles and pronouns, and did not separate the article from the verb with an apostrophe. As a consequence, to get an exhaustive list of occurrences of *enporter* in the AN texts, the query would have to match not only all the forms of the verb, but also all the possible kinds of agglutinated articles and pronouns, such as *d', s', m', l', n', c'*, etc. This is rather inconvenient and error-prone. Because of situations like these, we opted for "normalising" the texts prior to annotation. Normalisation has been previously applied to other historical languages, namely to Early Modern English (Rayson et al. 2007), Middle High German (Dipper 2010) and Early Modern German (Scheible et al. 2011) as well as to the ARCHER-texts (Hundt, Schneider, and Oppliger 2016), who all report a considerable increase of tagger-accuracy on normalized data (10%). In our case, recognition (i.e. the number of tokens matched in the tagger-lexicon) improves by 40 % on normalised

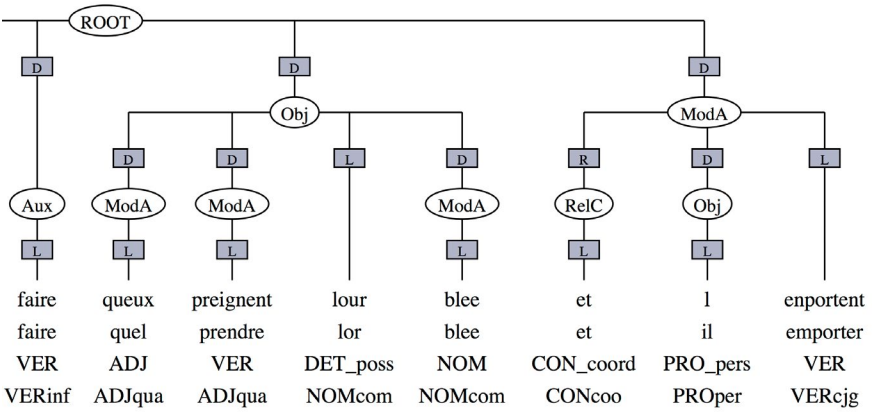

Figure 1: AN text data base, partial parse tree from albu783 (1225).

data. Unlike Dipper, and in line with Scheible et al., our approach does not involve retraining of the tagger. However, in contrast to Scheible et al, we do not intervene manually. Instead, we use an automated rule-based procedure and control the output of each single rule in order to prevent errors or over-generalisations. Also, in contrast to Rayson et al. and Scheible et al., we do not normalise to a modern standard and therefore do not have to intervene on the tagset itself, thereby maintaining straightforward comparability across both corpora.

Our goal is thus a POS-tagged and syntactically annotated version of the ANdb that allows us to retrieve, for example, not only the occurrence from the *Liber Albus* quoted in (3), just by searching for *enportent*, but ideally also the occurrence of "l" as a direct pronoun that is governed by the verb. The structure in Figure 1 is an example occurrence (see section 3 for an explanation of the dependency graphs).

Finally, we would like to point out that we did not normalise the text in the Lachmannian sense of the word. Rather, we calculated normalised forms in order to facilitate the identification of a given graphical form of the text in the tagger lexicon. If the generated form was successfully identified in the lexicon, the algorithm substituted the original form with the normalised form and did all further calculations on the basis of the generated form. But in the end, the generated form was, in turn, replaced by the original form, and all modifications that a given form had undergone remain invisible on the surface. In other words: no signs of intervention remain in the output.[7]

7 There is one change that nevertheless remains visible in the output, which is the separation of agglutinated forms. But this is state of the art in terms of "toilette du texte" (Foulet/Speer 1979, Lepage 2001 and École nationale des Chartes 2001). If a given text

2.2 Steps in normalisation

2.2.1 Preparatory measures

Of the 2,804,409 French tokens contained in the pre-processed version of the ANHub text-database, roughly two thirds (67.7 %) were matched by the Old French *TreeTagger* dictionary. Since the dictionary is all lower case, lower-casing all tokens raised successful identification to three quarters (75.09 %). At this point, we used a script to separate punctuation marks from words, because, given the fact that the dictionary contains only non-punctuated lemmata, tokens including punctuation marks would not have been retrieved. Tokenising increased the number of tokens to 3,439,145, four fifths of which were recognised (81.01 %).

The subsequent treatment described in the next sections below is based on this tokenised version of the ANdb (version 0.2). All the items that the tools could not identify in the dictionary at this point were submitted to further treatment.

2.2.2 Mechanical measures

In this step, we developed context-dependent rules for graphical normalisation. A graphical form that could not be matched in the first place underwent a series of successive context-sensitive modifications. However, while e.g. *lamour* is usefully converted into *l'amour*, *malaise* should be maintained as *malaise.* Therefore, each of these modifications was independently evaluated for success. This was easier with regard to pure graphical phenomena, such as e.g. the graphemes *y, k,* and *z*. In many cases, these graphemes are not used primarily for phonological reasons, but merely represent a variant spelling for *i, q(u)* and *(t)s* respectively. In these cases, they are fairly easy to replace, but the context has to be accounted for. E.g. *ey* equals *oi* in *neyent* 'nothing', *ai* in *faim* 'hunger', *eoi* in *receoit* 'he receives', *rey* in *derein* 'the last one', etc. In the end, y-rules were successfully applied in 23,164 cases.

The next step was to take into account regular phonetical features of AN, such as e.g. the spelling *ou* for *o*, or *om* for *ons*. These cases are of particular importance when it comes to suffixes, because, if a token such as *allom* 'we go, walk' was not recognised as a verb because of its ending in *om* instead of *ons*, the syntactic parser is also likely to fail at this point. Similarly, if *ioun* at the end of a word was converted into *-ion* and subsequently recognised as a word with a

lacks this kind of separation in the printed edition, it is no hallmark of the historical text, but the (modern) editors' choice. As such, there is no historical importance to it.

nominal suffix—that is, a noun—this benefits the part-of-speech and the syntactic analysis, even if the word is not in the tagger lexicon.

2.2.3 Additional measures

In addition to the substitutions described above, we had to intervene at two more points, one of them being proper names and the other agglutinated consonants. The latter keep the tagger from recognising the word even if the word itself is listed in the very same spelling in the dictionary, and in the case of the former, normalisation is not applicable.

Due to the nature of the texts included in the ANdb, many of which are legal documents and court proceedings, there are a considerable amount of proper names for both persons and places. In order to tag these adequately, we had to extract them and add them to the tagger-lexicon.[8]

We adopted two different approaches for extraction. Firstly, we collected all capitalised words from the file of unknown words generated by the tagger. In order to distinguish capitalised sentence initials from proper names, we sorted these forms by frequency, on the hypothesis that conjunctions etc., which might appear with a capitalised initial at the beginning of a sentence should occur as such more than once. In addition to that, sheer word-length allowed us to sort out a good deal of capitalised conjunctions, in contrast to proper names, which tend to be longer. This procedure allowed us to add 612 proper names to the tagger-lexicon and then re-train the tagger. Having again selected all capitalised forms from the new unknown-file, we sorted alphabetically, this time by the end of the words. This procedure helped us to detect the most frequent suffixes, such as e.g. *-fred* in person names or *-borough* and *-thorp* etc. in place-names. In the next step, we automatically extracted all forms ending in the 86 most frequent suffixes (down to frequency-rank 8) and added 2,473 additional proper names ending in the respective suffixes to the tagger lexicon. In total, we thus added 3085 additional entries. This step raised the overall recognition rate by roughly 1%.

The other approach dealt with agglutinated consonants. In AN, as in OF in general, words beginning with a vowel can combine with a consonant such as *c, d, l, m, n, q, qu, s, t* and the respective capital letters. A sequence like *l'article* would thus read *larticle* in the manuscript, and it would not be tagged correctly unless the agglutinated *l* was separated from the main word. On the other hand, this case has to be carefully distinguished from *malaise*, which should not be split into *m'alaise*.

8 Most taggers also exploit the „suffixes" of words to predict the category; the TreeTagger also applies such an algorithm to unknown word forms.

Therefore, we included a routine that checks unknown words for the initial sequence of "agglutinate consonant + vowel". If a word matches this pattern, the algorithm experimentally splits off the consonant and resubmits both elements to the recognition-procedure, this time analysing both parts independently, and writing successfully treated forms into an extra file. This output was checked manually. False recognitions, such as *d'estrece* 'narrowness' built from *destrece* 'hardship, affliction' (hypothetical example) were collected in a separate file. The routine then checked this file before proceeding to the treatment of possibly agglutinated forms. Doing this, the number of tokens was raised to 3 448 633, and, based on this new number, the rate of forms recognized by the tagger is at 92.94 %.

As one can see, it is indeed possible, and even at very low cost, to raise the rate of recognition by some 40%. One way to achieve this is by preprocessing the texts through "normalisation". By applying the procedures described above, we were able to normalise about 164,000 tokens equalling 39 000 types, with maximum token frequencies of up to 1,340 for forms of *estre* 'be' (1,340 *sunt*, normalised to *sont*, 3.pl.ind., and 1,328 *seyt*, normalised to *soit*, 3.sg.subj.).

The other way to raise recognition is by adapting the tagger and its lexicon, as they had originally been trained on continental French data, in order to cope with AN texts. Overall, "normalization" increased the rate of AN forms that are successfully identified in an OF tagger-lexicon by 25 percent points, from 67.7 % to 92.94 %—a step which will be crucial for the subsequent syntactic analysis.

3 Automatic syntactic analysis

3.1 Old French corpus annotation applied to Anglo-Norman

After the "normalisation" of the data described in the previous section, we applied a part-of-speech tagger and a dependency parser to the ANdb. Both were previously trained on Old French text corpora.

For part-of-speech annotation and lemmatisation, we used the *TreeTagger* with parameters for Old French. The tagger was trained on the *Nouveau Corpus d'Amsterdam* (Kunstmann/Stein 2007) and used a lexicon with form-tag-lemma triples that were extracted from various Old French resources[9]. This lexicon was identical to the one that was used for verifying the output of the normalisation rules described in section 2.3.

9 The training of *TreeTagger* and the lexical resources are described in Stein (2007). The lexical resources are freely available as *FROLEX*, see https://github.com/sheiden/Medieval-French-Language-Toolkit.

For dependency annotation we decided to use the *mate tools*[10] *joint transition-based parser* (Bohnet et al. 2013) for joint part-of-speech tagging and parsing. The parser was trained on the dependency annotation of the *Syntactic Reference Corpus of Medieval French* (SRCMF, Prévost/Stein 2013). The training corpus extracted from SRCMF contained 12 texts or text samples, written between 1000 and 1300, and containing 242,946 word tokens (23,818 types). Punctuation was not present (since modern punctuation appears only in modern transcriptions), and orthographical variation was considerable: the type-token ratio was more than twice as high (0.099) than in average Modern French texts (0.05), with the obvious negative consequences for the precision of part-of-speech tagging. The syntactic categories in the training corpus were a slightly simplified set of the SRCMF categories (see the documentation on the corpus web site http://srcmf.org).

The *joint transition-based parser* was chosen because it performed slightly better than the *mate tools* graph-based parser (Bohnet 2010) we had trained on the same corpus. Accuracy scores were better both for part-of-speech tags and labeled dependency attachment. More importantly, the joint transition-based parser also attained a higher score of exact sentence matches (i.e. where all the dependencies and categories in a sentence were analysed correctly) on our Old French evaluation corpus. The training procedure and the two *mate tools* parsers are described in greater detail in Stein (2016).

Concerning the results of this parser as applied to the Anglo-Norman texts, our expectations are not high. With a labeled attachment score of 85.96 % and a score of 47.59 % for exact sentence matches on the evaluation part of the SRCMF (i.e. a corpus containing the same text types), it is clear that the uncorrected output will present a considerable number of errors. Due to the particular characteristics and the heterogeneity of the AN texts described above, the parser is bound to perform worse, and we expect only very short sentences to be parsed correctly. An example for such a short sentence with correct analysis is given in (4), where according to the SRCMF markup, "Cmpl" is the indirect object, "RelNC" a non-coordinating relator (here: preposition), "Obj" the direct object, and "ModA" a modifier (including also determiners):

(4) A lui comand la meie vie
To him command.1.SG the my life
'I command my life to him.' (1250resu)

The output format of the parser is the CoNLL 2009 tabular format (defined on the CoNLL 2009 shared task web site, see http://www.conll.org). For the sake of clarity, Figure 2 shows a simplified CoNLL format representing only selected

10 https://code.google.com/archive/p/mate-tools/

columns: word number, form, lemma(s), TreeTagger POS, parser POS, morphological features, head attachment, and dependency relation. The last two columns encode the dependency structure. For example, "0" marks the verb *comand* as being the root node. "3" attaches *lui* (word no. 2) to *comand* (word no. 3), and the dependency relation is "Cmpl", i.e. indirect object. Likewise, "2" attaches *A* (word no. 1) to *lui* as a "RelNC" (non-coordinating relator), and so forth.

```
1  A       a          PRE       PRE        _                          2  RelNC
2  lui     il|loi     PRO:pers  PROper     G=masc|N=sg|C=obj|P=3      3  Cmpl
3  comand  comander   VER       VERcjg     N=sg|P=3                   0  ROOT
4  la      le         DET:def   DETdef     G=femi|N=sg|C=obj          6  ModA
5  meie    mien       ADJ:poss  ADJqua     G=femi|N=sg|C=obj          6  ModA
6  vie     vie|voie   NOM       NOMcom     G=femi|N=sg|C=obj          3  Obj
```

Figure 2: CoNLL format (simplified) for sentence (4).

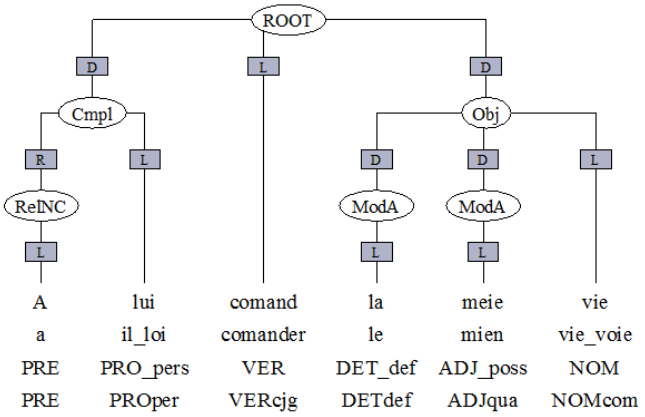

Figure 3: TigerSearch graph for sentence (4).

The CoNLL format can be used directly with some query tools like *Icarus* (Gärtner 2010). However, in the next section we use *TigerSearch* queries (Lezius 2002), since this is the default distribution format of the Old French SRCMF corpus. We therefore converted the CoNLL output of the parser into TigerXML. In Figure 3, the structure is shown as represented in the *TigerSearch tool*. In order to represent the SRCMF dependency graphs in *TigerSearch* (which was primarily designed to represent constituency structures), we distinguish between two kinds of relations (arcs): the default relation is dependency, labelled with a "D", whereas "L" marks the unique lexeme that governs the structure and

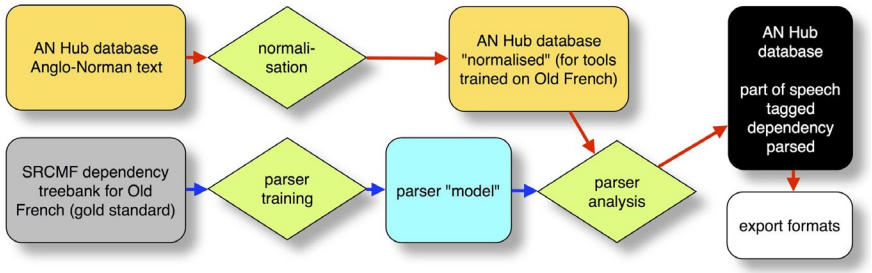

Figure 4: Annotation of the Anglo-Norman text database.

would figure as the top node of the structure in a traditional dependency graph à la Tesnière. For example, the main verb *comand* is attached to the root node by the "L" relation.

The complete workflow of the annotation is resumed in the flow chart shown in Figure 4. In the next section we will discuss the usability of the output.

3.2 Usability of unsupervised parsing

3.2.1 A case study

Since there is no gold standard corpus for AN, we cannot provide a quantitative assessment of the annotated ANdb. The goal of the cooperation between the BASICS and the AND projects was meant to be a feasibility study rather than an annotation project in its own right. We decided to use the annotated output for a research question that was relevant for the BASICS project anyway: the variation between direct and indirect objects that was observed in AN e.g. by Ingham (2010). These cases of variation are, for example, relevant for the development of passive structures in the medieval contact situation between English and (Anglo-)French. As pointed out in Stein and Trips (accepted), language contact with OF and AN may have attributed to the rise of the recipient passive (e.g., in ModE, *She was given the book*), since in Middle English corpora, the first occurrences of the recipient passive appear predominantly with verbs of French origin. So our analysis bears on OF ditransitive constructions. Just as in Modern French, continental OF had a dative goal (or recipient) phrase, i.e. a prepositional phrase governed by *a* (ModF à), for example with the verb *demander* ‚ask', as in sentence (5):

(5) et demande a Lancelot quele aventure l' a ilec amené
and asks.3.SG to Lancelot which adventure him has here brought
'and asks Lancelot which adventure brought him here (SRCMF, qgraal)'

One of the hypotheses we wanted to verify using the annotated ANdb was that the argument structures of AN ditransitive verbs was different from the (continental) OF structures, showing variation between indirect and direct objects. In order to do so, we needed to extract these verbs in specific constructions from the corpus. In the following subsections, we describe the relevant queries step by step, from the word level to the syntactic level, and discuss the advantages and problems we encountered in the annotation.

3.2.2 Lemmatisation

At word level, the first step was the selection of a representative sample of clause-taking verbs. We used the lemmatisation introduced by *TreeTagger* to query for eight such verbs, i.e. *assëurer, demander, certefiier, comander, garnier, informer, prier, vëer.*[11] We manually checked the precision of the results. It was generally satisfying, i.e. the result did not contain many forms not matching these verbs, except for some prefixed forms (forms of *deprier* instead of *prier*). The recall (i.e. the relation between the extracted forms and those which could have been maximally extracted) can only be estimated. We again verified manually and found that recall was not lower than if we had performed a search targeted at inflected forms, using regular expressions. This is probably due to the fact that AN graphical variants are fairly unpredictable (as was shown in section 2). Nevertheless, by querying the lemmas we found a number of graphical forms that would have been hard to guess, as for example *Nos te praeiam* (*nous te prions*, 1.PL., 'we pray you'). And queries aiming at particular verb classes (which often have many more than the eight members we selected for our example) would be extremely laborious if lemmatisation was not present in the annotation. So we can conclude that unsupervised lemmatisation, even if it is only partial and may contain errors, is indeed useful.

11 The TigerSearch query specified the following lemmas:
`[lemma=/.*(assëurer|demander|certefiier|comander|garnier|informer|prier|vëer).*/]`

3.2.3 Ditransitive constructions

The second task was to narrow the output down to ditransitive constructions. This step *requires* syntactic annotation (or manual analysis, which is not at issue here). Querying ditransitive constructions using only part-of-speech annotation is extremely laborious. It requires a combination of several subsequent queries, and would probably lead to low precision and recall values. This is due to the variable position of each of the arguments in the clause, the graphical variants, and the syntactic ambiguities, where the first two factors affect recall, and the latter affects precision (not every prepositional phrase is an indirect object, etc.). The SRCMF grammar model reproduced by the parser allows extraction of ditransitive constructions with a single query, e.g. in *TigerSearch*. This query[12] finds a total of 365 sentences (265 with a verb form of *demander*, 168 with *comander*, 26 with *prier*, 26 with *vëer*, etc.).

3.2.4 Clitics vs full lexical arguments

In the next step, we were interested in the various forms of argument realisation. Since AN texts sometimes show inconsistencies in the use of clitics (Old French distinguishes between accusative and dative pronouns), we are interested in the different combinations of clitic and full lexical argument realisation. Clitics appear preverbally, i.e. at a position that is normally different from the (generally postverbal) position of full nominal arguments. Again, clitics are very difficult to retrieve unambiguously: their graphical forms are extremely variable, and they are often homographs of other grammatical morphemes like articles (*le*, *li*, etc.). Using the syntactic annotation, we retrieved clitics by combining POS tag ("PRO") and node "arity", the latter being "1", since clitics do not govern other nodes. The *TigerSearch* query given in footnote[13] is meant to serve as an example: it extracts only the occurrences where both the direct and indirect object are clitics, which is in fact quite rare. In our project, we are rather interested in cases that are analogous to *She commanded him to leave*, in order to find out if the goal argument (*him*) is an accusative or a dative clitic in the Anglo-Norman construction. So, one of the arguments needs to be specified as being clausal. We

12
```
#s:[type=/V.*/]
   & #s > #v:[< list of lemmas, as needed>]
   & #s > #a1:[cat="Obj"]
   & #s > #a2:[cat="Cmpl"]
```

13
```
[lines 1-4 identical to first query]
   & arity(#a1,1) & #a1 > #acc:[pos=/PRO.*/]
   & arity(#a2,1) & #a2 > #dat:[pos=/PRO.*/]
```

further restrict our clitic query to third-person forms beginning with *l* (since first and second person do not distinguish between accusative and dative). Finally, the goal argument, which in continental Old French normally has dative case, is specified as direct object ("Obj", i.e. accusative).[14]

We applied this query to the SRCMF corpus as well as to the ANdb. In SRCMF we obtained only one result (from the *Chanson de Roland*, an early Anglo-Norman text):

(6) Par penitence les cumandet a ferir
By regret them.ACC commanded to strike
'He regretfully commanded them to strike.' (roland-pb: 100-lb1138)

In the ANdb, the query retrieved ten occurrences, which could confirm that the variation between dative and accusative clitics in clause-taking ditransitives is indeed characteristic of Anglo-Norman. The precision, however, was low: in addition to the ten valid examples we retrieved many erroneous hits where the parser annotated the wrong structure. A typical error is the non-recognition of dislocations, as in example (7):

(7) donets moy grace qe jeo le voille et jeo soie si treshumble pacient come
le mestier le demande a receivoir bonement les cures
the profession$_i$ it.ACC$_i$ requires [to accept well the cures]$_i$
'The profession requires it to receive the treatments willingly.' (1354seyn2374)

Even for the human reader, it is not an easy task to detect that *le* preceding *demande* is not the goal argument here, but a cataphoric clitic that doubles the right-dislocated clausal complement, i.e. *a receivoir bonement les cures* (both are co-indexed with *i* in the glossed example). So in fact, this example is not an instance of ditransitive *demander*. Again, if we wanted to measure the recall of the query we would have to check for missed occurrences, using a series of word form and POS-based queries.

The last variant we discuss here is the case where the goal argument is a full NP and the clause is the theme. Again we want to find out if the goal argument is a direct or indirect (prepositional) argument, i.e. "Obj" or "Cmpl" in terms of SRCMF categories (analogous to English constructions like *She commanded (the*

14
```
#s:[type=/V.*/]
   & #s > #v:[<list of lemmas, as needed>/]
   & #s > #a1:[cat=/Obj|Cmpl/ & type=/V.*/]
   & #s > #a2:[cat="Obj"]
   & arity(#a2,1) & #a2 > #dat:[word=/l.*/ & pos=/PRO.*/]
```

knight|to the knight) to leave). In the query given in the footnote[15], we defined the goal argument as non-verbal, specifying a minimal arity of 2 (thus eliminating clitics) and added a restriction for linear precedence (goal occurring before clause). Again, precision was low: the query produced noise due to parsing errors in complex sentences. A good result is example (8), whereas in example (9) the subject was wrongly parsed as a direct object:

(8) ... et demandent **les marchans** a avoir du maistre leurs denrees
... and ask.3.PL the merchants.ACC to have from-the master their goods
'and they ask the merchants to get their goods from the master'
(1310domg1769)

(9) Et comande le Rei qe les Viscontes ...
And comands the king.NOM that the viscounts ...
'and the king commands that the viscounts ...' (1275stat110)

3.2.5 Analysing grammatical variation

A particular problem arises when the corpus analysis targets grammatical variation. Variations like the one mentioned above, between accusative and dative clitics, are notoriously difficult to identify using machine-learning approaches. In our case, the variation is said to be typical for later AN. Since the parser was trained on the SRCMF texts, it cannot be expected to have encountered this kind of variation. Therefore, when the less frequent option of a particular instance of grammatical variation is encountered in the input data, this will create a conflict at the syntactic level. In the example (6), the clitic *les* is part-of-speech tagged as accusative, but it co-occurs with a verb that normally governs a dative complement (*comander*). It is rather unpredictable, at least for the linguistic user, if the parser will select the category, i.e. direct *vs* indirect object, that matches best the part-of-speech analysis or the valency of the verb. In our corpus, the *joint transition-based parser* seemed to be more strongly influenced by the part-of-speech information. That means that the linguistic perspective, which describes this case as variation on the morphological level, cannot be translated directly into a

15
```
#s:[type=/V.*/]
& #s > #v:[< list of lemmas, as needed>]
& #s > #a1:[cat=/Obj|Cmpl/ & type=/V.*/]
& #s > #a2:[cat="Obj" & type!=/V.*/]
& arity(#a2,2,99) & #a2 .* #a1
```

query. Instead, the user has to anticipate the way the parser analyses these cases when formulating their query. Examples like (6) can only be retrieved by a query that specifies the goal argument as direct object ("Obj") on the syntactic level or underspecifies the syntactic category.

4 Conclusion

The goal of this contribution was to demonstrate how linguistic tools that were previously trained on other varieties of a medieval language can be applied to a specific variety of this language using „normalisation" techniques. In our case, the medieval language was Old French (OF), and the new corpus was the Anglo-Norman text database (ANdb). Since graphical conventions in Anglo-Norman (AN) are quite different from those of continental OF, we normalized the AN texts before applying the computational-linguistic tools. We use "normalising" in the sense of adapting the AN forms to the continental OF spelling conventions as closely as possible. We used the OF lexicon contained in the parameters of *TreeTagger* to measure the score of normalised forms and showed how graphical normalisation, including the resolution of determiners that are agglutinated to nouns, improves the performance of the tools. We (partly) lemmatised the corpus using *TreeTagger*, and added dependency structures using the *mate-tools joint transition-based parser*. Since a gold standard corpus for Anglo-Norman does not exist, we were unable to calculate accuracy scores for these analyses. Instead, we evaluated the quality of the annotation from a linguistic point of view, searching for particular argument realisations of ditransitive verbs.

As expected, the major issue due to errors in the annotated version of the ANdb is low recall, and it is hardly measurable how many of the structures we queried were not successfully retrieved. We showed that, in some cases, a good feeling for the way the parser works is required to anticipate its analyses and to formulate the queries accordingly. This issue hampers the quantitative interpretation of the data. However, we also saw that parsing, albeit imperfect, allows us to make queries and extract occurrences for structures we could not have retrieved otherwise (at least not in acceptable time). Thus, even with medieval texts, the unsupervised use of computational tools, paired with a normalisation procedure that graphically adapts the novel text to the graphical conventions of the training corpus can help to extract relevant syntactic data and thus assist diachronic syntactic analysis. Especially with larger amounts of data (as in the case of the ANdb, containing over 3 million words) parsing, even with low accuracy, may be the only way to discover certain phenomena and to retrieve the relevant data.

References

Bohnet, Bernd. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China: Coling 2010 Organizing Committee, 89–97.

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter and Jan Hajic. 2013. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *TACL 1*, 415–428.

Dipper, Stefanie. 2010. POS-tagging of historical language data: First experiments. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10)*, Saarbrücken.

Gärtner, Markus, Gregor Thiele, Wolfgang Seeker, Anders Björkelund and Jonas Kuhn. 2013. ICARUS – An Extensible Graphical Search Tool for Dependency Treebanks. *Proceedings of ACL 2013*.

Grant, Judith. 1978. *La passiun de seint Edmund.* Oxford: Blackwell.

Hundt, Marianne, Gerold Schneider, Rahel Oppliger 2016: Part-of-Speech in Historical Corpora: Tagger Evaluation and Ensemble Systems on ARCHER. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS-13)*, Bochum.

Hunt, Tony. 2004. *Le Chant des chanz.* London: Anglo-Norman Text Society.

Ingham, Richard. 2010. The Transmission of Later Anlo-Norman: Some Syntactic Evidence. In Richard Ingham (ed.), *The Anglo-Norman Language and its Contexts*, 164–182. Woodbridge: Boydell and Brewer.

Ingham, Richard. 2012. Middle English and Anglo-Norman in Contact. *Bulletin de l'Association des Médiévistes Anglicistes de l'Enseignement Supérieur* 81: 1–23.

Johnston, Ronald Carlyle. 1961. *Crusade and Death of Richard I.* Oxford: Blackwell.

Kunstmann, Pierre and Achim Stein. 2007. Le Nouveau Corpus d'Amsterdam. In Pierre Kunstmann and Achim Stein (eds.), *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, 9–27. Stuttgart: Steiner.

Lezius, Wolfgang. 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora (German).* Stuttgart: Institut für Maschinelle Sprachverarbeitung (IMS).

Marchello-Nizia, Christiane. 2009. Histoire interne du français: morphosyntaxe et syntaxe. In Gerhard Ernst, Martin-Dietrich Gleßgen, Christian Schmitt and Wolfgang Schweickard (ed.), *Romanische Sprachgeschichte. Ein internationales Handbuch zur Geschichte der romanischen Sprachen und ihrer Erforschung, Teilband 3*, 2926–2947. Berlin, New York: de Gruyter.

Prévost, Sophie and Achim Stein. 2013. *Syntactic Reference Corpus of Medieval French (SRCMF).* Lyon/Stuttgart: ENS de Lyon; Lattice, Paris; Universität Stuttgart.

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS

tagger on Early Modern English corpora. In *Proceedings of the Corpus Linguistics Conference 2007*. Birmingham: University of Birmingham.

Rothwell, William and David Trotter. 2005. *Anglo-Norman Dictionary 2. Online Version.* London: MHR.

Scheible, Silke, Richard J. Whitt, Martin Durrell and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 19–23.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. In Daniel Jones (ed.), *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP'94), Manchester, September 1994*, 44–49. Manchester: UMIST.

Stein, Achim. 2014. Parsing Heterogeneous Corpora with a Rich Dependency Grammar. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 26.–31.5.2014*, Reykjavik, Iceland: European Language Resources Association (ELRA).

Stein, Achim. 2016. Old French Dependency Parsing: Results of Two Parsers Analysed from a Linguistic Point of View. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 23.–28.5.2016*, Portoroz, Slovenia: European Language Resources Association (ELRA).

Stein, Achim, Carola Trips (accepted): A comparison of multi-genre and single-genre corpora in the context of contact-induced change. In Richard Whitt: *Diachronic corpora, genre and language change*. Amsterdam, Philadelphia: Benjamins.