Achim Stein (Stuttgart) / Martin-D. Gleßgen (Zürich)[1]

Resources and Tools for Analyzing Old French Texts

Cette contribution décrit les travaux visant à l'étiquetage et la lemmatisation automatique de textes en ancien français. Plusieurs ressources lexicales et textuelles ont été réunies pour dresser un lexique de formes fléchies, qui, à son tour, est utilisé pour étiqueter et lemmatiser automatiquement des textes. Après entraînement sur un corpus pré-annoté manuellement, un étiqueteur probabiliste attribue à environ 96% des mots d'un texte la partie de discours correcte et propose un ou plusieurs lemmes pour environ 57% du vocabulaire. Ce résultat peut être amélioré par une routine semi-automatique complémentaire.

1. Presentation of the project

The project *Sources et outils pour l'analyse du français ancien* (SOFA) is a collaboration between the Laboratoire de Français Ancien (LFA, University of Ottawa, Canada), the Institut für Linguistik/Romanistik (ILR, University of Stuttgart, Germany), the Romanisches Seminar (RoSe, University of Zurich, Switzerland), and other associated researchers.[2] The LFA contributed its lemmatized base of Old French texts and several lexical resources mentioned below. The ILR provides tools and expertise in the field of automatic morphological and syntactic analysis. The overall goal is to bring together technology and resources for the automatic treatment of Old French on a reliable and modern philological basis.

2. Resources and tools

2.1. Lexical resources

[1] The paragraphs 2.4 and 2.6 are written by M.-D. Gleßgen.

[2] The collaboration between Ottawa and Stuttgart is funded by the Conseil de recherches en sciences humaines du Canada (CRSH), the University of Ottawa (Faculty of Arts) and the Alexander von Humboldt-Stiftung (Transcoop Programme).

1. The electronic version of the Tobler/Lommatzsch *Altfranzösisches Wörterbuch* (Blumenthal/Stein 2002) contains a list of all the 37.000 lemmas of the dictionary as well as 15.000 cross-references, which are mostly spelling variants of the lemmas. By courtesy of the publishing house, Franz Steiner Verlag, Stuttgart, these lemma lists and other materials are publically available.[3]
2. The base of verb forms: In the 1960s Robert Martin extracted the verb forms of the most important historical dictionaries and manuals for Old and Middle French up to the 16th century (Tobler/Lommatzsch, Godefroy, Huguet) as well as some critical editions. Each record contains the verb form, morphological information and the lemma. In a joint project between the former Institut National de la Langue Française (formerly INaLF, now ATILF) and the Laboratoire de Français Ancien (Ottawa), these records were transformed into a database and published on the LFA web site. The verb database contains about 56.600 forms and has added 37.400 new forms to our lexicon (19.200 forms were already in the Amsterdam Corpus).
3. Following the example of Martin, graphical forms were also extracted from all the articles of the Godefroy dictionary at the LFA. These forms are the lemmas, the variants cited after the lemmas, and the inflected forms which occur in the examples. At the time being (February 2004), we have included the forms of the articles AA-AN, EM-L, M-OK and T-Z. These 130.000 forms are annotated with part of speech tags (however without any further morphological information) and the lemma.
4. In the next step, we will integrate lemmas and inflected forms attributed semi-automatically to the Old French texts edited by the LFA and to the different volumes of the *Plus anciens document linguistiques de la France*[4] (see below).

2.2 Text resources

1. The *Amsterdam Corpus of Literary Texts* (AC) was compiled in the beginning of the 1980s by a group of scholars directed by Anthonij Dees and resulted in the *Atlas des formes linguistiques des textes littéraires de l'ancien français* (Dees 1987). The electronic version of the AC is provided by Piet van Reenen (Free University of Amsterdam). It contains about 200 different texts, some of them in several manuscripts, which adds to a total of 289 text samples with about three million words (tokens). These forms have been manually annotated by Dees' team with a set of 225 numeric tags encoding part of speech and other morphological categories (e.g. "566" for verb, futur tense, 3rd person, plural). Some of the texts are electronic versions of existing editions (e.g. the *Miracles de Notre Dame de Chartres* by Jean le Marchant,

[3] http://www.uni-stuttgart.de/lingrom/stein/tl/
[4] This project is directed by Françoise Vielliard, Olivier Guyotjeannin and, for the lexicological and informatical part, M.-D. Gleßgen.

edited by P. Kunstmann, Chartres/Ottawa, 1973), others are transcriptions of manuscripts made especially for this corpus. Despite of its fine grained morphological markup, the AC has not been lemmatized. Nevertheless it is a precious resource which enables us to extract a lexicon of more than 130.000 Old French inflected forms, and, what is more, to train the TreeTagger (see below).

2. Several Old French Texts have been published by Pierre Kunstmann on the web site of the Laboratoire de Français Ancien, University of Ottawa. We exploited the indices of two texts: *Le conte du Graal* by Chrétien de Troyes (edited by P. Kunstmann) and *Le couronnement de Louis* (edited by Y. Lepage). For each lemma (taken from the Tobler/Lommatzsch), the index indicates its part of speech and an ordered set of all inflected forms, each set consisting of the number of occurrences, the lemma and the list of references.

3. The collection of the *Plus anciens document linguistiques de la France* provides a lemmatization for all lexical words, realized by means of the tool *Phoenix* (see Gleßgen 2003a; Matthey in this volume; and 2.6 below); the existent sets of inflected forms can be integrated in the all over routine.

4. A further resource, which has been compiled manually, is the inventory of grammatical morphemes. They represent a particular problem, because the Amsterdam Corpus does not distinguish between certain categories (probably because they were not of interest for Dees' work): for example the tag "600", marks ambiguous graphical forms which can be adverbs, conjunctions or pronouns (*ce, ne, que, qui, ou* and their variants), and the form *mais* is always marked as a conjunction irrespectively of its potential adverbial sense. Such cases are problematic in two ways: first, because they will have to be disambiguated manually in a revised version of the corpus, and second, because they provide the wrong distributional input for the training of the TreeTagger. For the time being we decided not to correct the manual markup of the AC but to focus on lemmatization: about 4.000 grammatical morphemes were extracted from the AC, revised and associated with 134 Tobler/Lommatzsch lemmas. In the final markup these lemmas are marked with "_S" (see table 2). The assignment of forms to categories follows, if possible, the CATTEX conventions established for the markup of the *Base de francais médiéval* (Chr. Marchello-Nizia, ENS-LSH Lyon, see Heiden/Prevost in this volume), although the abbreviations are not the same. The 134 lemmas correspond to the part of speech tags listed in (1), where the number of forms and an example are given for each category.

(1) ADJ:poss (438, e.g. *mien*), CON:coord (23, e.g. *car*), DET:def (57, e.g. *li*), DET:demo (177, e.g. *cel*), DET:ind (646, e.g. *alcun*), DET:ndf (79, e.g. *un*), DET:poss (401, e.g. *nostre*), PRE (448, e.g. *auec*), PREDET:a (47, e.g. *al*), PREDET:de (41, e.g. *del*), PREDET:en (56, e.g. *el*), PRO:clit (23, e.g. *en*), PRO:demo (341, e.g. *cela*), PRO:ind (430, e.g. *alcun*), PRO:invar (247, e.g. *quoi*), PRO:pers (310, e.g. *el*), PRO:poss (238, e.g. *mien*).

2.3. Merging the lexical resources

The resources described above are merged into a "lexicon of Old French forms". A filter programme converts the morphological information into a standardised format required by the TreeTagger and creates a uniform set of 50 tags. The resulting lexicon totals 336.500 forms (of which 252.300 are graphically different).

The tags distinguish the part of speech and some minor categories: subtypes of adjectives (e.g. numerals) and pronouns (e.g. indefinite, interrogative). Although most of our resources provide more information (e.g. person, gender, number, verb tense), these categories are missing in the indices and in the Godefroy database. Since the definition of a complete final tagset is not an issue at this stage of the project, we reduced the tagset to the minimal information shared by all the resources. For philological reasons, it would of course be desirable to extend all the tags to the most explicit one at a later stage, for example to the tagset of the Amsterdam Corpus.

Merging the resources does not modify the lemmas they provide: if the lemmas differ for a given form, they are listed as alternatives in the lexicon. Uppercase letters preceded by an underscore indicate the resource which has provided the form:

- "G" for the Godefroy dictionary,
- "I" for the LFA texts (indexes),
- "M" for the verb forms compiled by Robert Martin,
- "S" for the list of grammatical morphemes,
- "T" for the Tobler/Lommatzsch lemma list.

A programme which uses "morphological" rules suggests a lemma for the 127.000 unlemmatized forms in the lexicon (these are the forms from the AC which did not match with lemmatized forms from other resources). At present, about 50 rules deal with the most frequent endings of unlemmatized forms by simply stripping off a predefined string and optionally adding another string. If the result matches an existing lemma with the same part of speech tag, the lemma is adopted for the unlemmatized form, but marked with an asterisk as being "constructed".

abitable	ADJ	<nolem>		
abitablement	NIL	habitablement_G		
abitacion	NOM	+abitacion_T		
abitacle	NOM	+abitacle_T		
abitance	NIL	habitance_G	NOM	+abitance_T
abitant	NOM	+abitant_T	VER	*+abiter_IT
abitanz	NOM	*+abitant_T		
abitanze	NIL	habitance_G		
abitast	VER	<nolem>		
abitateur	NIL	habitateur_G		
abitation	NOM	*+abitacion_T		
abitations	NOM	*+abitacion_T		
abitator	NOM	+abitator_T		
abite	VER	+abiter_I		
abitee	NIL	habiter_G	VER	*+abiter_IT

Table 1. The lexicon of Old French forms

The final step is the comparison of each lemma with the list of the Tobler-Lommatzsch lemmas: Each lemma that appears in the Tobler-Lommatzsch is marked with a plus sign.[5]
Table 1 shows some sample entries from the resulting lexicon. The first column contains the graphical form, the following pairs of columns contain the part of speech tag and the lemma. Ambiguous forms (like *abitant* in the example) have more than one tag-lemma combination. Some forms like *abitablement* have no part of speech tag (hence "NIL"), others like *abitable* have no lemma. For *abitance* two different lemmas are provided (from the Godefroy and the Tobler/Lommatzsch respectively), and for *abitanz, abitation(s),* and *abitee,* the "morphological" rules suggested a matching lemma (*), which could also be verified in the Tobler/Lommatzsch lemma list (+).

2.4. Philological aspects and historical linguistics

Parallel to the development of the markup procedure, the AC is submitted to a philological review on the basis of the standards of the other text resources. The description of the diasystematic and philological parameters is specified for the 289 text samples (dating and localisation of the manuscripts and texts; principles and quality of the edition). A certain number of editions are not reliable and might be excluded from the corpus.
On the other hand, the consideration of a large amount of forms will lead to the consolidation of the morphological tagset and the set of lemmata for the *langue d'oïl.* The coexistence of forms from different regional *scriptae,* different genera and different periods induces a high degree of complexity in

[5] Due to its particularly well conceived construction, the Tobler/Lommatzsch is still a reference for many scholars, at least as long as the *Dictionnaire étymologique de l'ancien français* (DEAF) is not completed.

the lemma definition. The reference set of the Tobler-Lommatzsch lemmas will thus be enlarged, to enhance the knowledge about the lexical structures of the older stages of French (cf. Gleßgen 2003b).

2.5. Automatic part of speech tagging and lemmatisation

The TreeTagger is a probabilistic part of speech tagger which uses decision trees. It was developed by H. Schmid (IMS, University of Stuttgart). Contrary to other probabilistic tagging methods, which have difficulties in estimating small probabilities accurately from limited amounts of training data, the TreeTagger avoids the sparse data problem by using a binary decision tree which determines the appropriate size of the context used to estimate the transition probabilities. Possible contexts are not only trigrams, bigrams and unigrams, but also other kinds of contexts (e.g. tag_{-1}=ADJ and tag_{-2}≠ADJ and tag_{-2}≠DET, for technical details see Schmid 1994). During the lookup of a word in the lexicon of the TreeTagger, the lexicon is searched first. If the word is found there, the corresponding tag probability vector is returned. If not, the TreeTagger tries to guess the right tag from the last letters of the word (suffix probabilities). So far TreeTagger modules (parameter files) have been developed for English, German, Modern French (Stein/Schmid 1995) and Italian.[6] The TreeTagger consists of two separate programmes for training and tagging. The input for the training consists of several files: the lexicon, as described above, some minor files containing the tags for open classes (i.e. the categories for which a suffix decision tree is built in order to guess the category of unknown words), and the training text. For the training, the Amsterdam Corpus was split up in two parts: the larger part (about 2.6 million words) was used for the training, the smaller part (500.000 words) was used for the evaluation of the annotation. The output of the training is a single parameter file which contains the lexicon and the decision tree data. Only this parameter file and the tagger binary are required to annotate new texts.
Table 2 shows some lines of disambiguated output[7], one word per line, with the part of speech tag in the second and the lemma(s) in the third column. The TreeTagger selects the tag with the highest probablity and inserted the corresponding lemma (it is of course possible to prevent the tagger from taking this decision and to display instead all the possibilities with their respective probabilities. In this case, the annotation has to be disambiguated manually). All lemma forms are shown, graphical variants (e.g. *mostier* vs. *moustier*) as well as different solutions for a given form (e.g. *tens, tant, taon*).

---

6 The TreeTagger and these parameter files are freely available for Linux, Solaris, and Mac OS-X (see WWW address given at the end of this contribution).

7 From *L'histoire de Barlaam et Josaphat,* as included in the AC.

an	PRE	+en1_IS|en_I
cel	DET:demo	cel_S
tans	NOM	+tens_I|tant_G|taon_G|+tens_G
que	PROCON	<nolem>
li	DET:def	le_ST
mostier	NOM	+mostier_IT|moustier_G
et	CON:coord	et_ST
les	DET:def	le_ST
yglises	NOM	*+eglise_IT
furent	VER	+estre1_I|estre_MI
conmancié	VER	<nolem>
a	PRE	+a3_T|a_GIS
edifier	VER	<nolem>
ou	PROCON	o_G|od_G
non	ADV	+non_G
nostre	DET:poss	nostre_SG
signor	NOM	+seignor_T
jhesu	NPR	Jhesu_I|<nolem>

Table 2. TreeTagger output

In view of the fact that the resources are incomplete, the results are encouraging. In the evaluation corpus, 83.4% of the tokens and 56.6% of the types have been lemmatized. The precision of the morphological markup is close to 96.3%, the most frequent error being the confusion of *en* preposition with *en* conjunction (see table 3). Other errors reveal shortcomings in the manual annotation of the Amsterdam Corpus.

Errors	**Form**	**manually assigned**	**TreeTagger**
2011	*en*	PRO:clit	PRE
976	*ne*	PRO:clit	PROCON
722	*a*	VER	PRE
634	*ne*	PROCON	PRO:clit
351	*i*	PRO:pers	PRO:clit
331	*en*	PRE	PRO:clit
328	*a*	PRE	VER
310	*de*	NOM	PRE
188	*c*	PRO:invar	PROCON
184	*n*	PROCON	PRO:clit

Table 3. The most frequent errors

In order to evaluate the coverage of our lexicon of Old French forms, we calculated the percentage of unknown forms in some texts which are not included in the AC nor have they been exploited as direct resource for our lexicon. For the samples between 1100 and 1300, the results were mostly below 10%. For the *Miracles de Nostre Dame* by Geoffrey de Coinci, who is well known for his neologisms, we got results ranging from 6.6% (vol. 1) to 21.5% (vol. 4). As expected, the percentage of unknown forms increased in more recent texts and went up to over 43% for Froissart's *Chronique* (about 1400).

2.6. Semi-automatic desambiguation of lexical words

The tree tagger provides the probability estimated for the markup of any form (with the precision indicated in 2.5). The ambiguous forms need further examination: If for highly frequent grammatical words probabilistic methods seem appropriate, for lexical words an automatically sustained manual procedure is more adequate.
To these means, the tool Phoenix has been developed since 2000 by M.-D. Gleßgen and Matthias Kopp, University of Tübingen. The module adapted by Matthias Kopp selects all ambiguous forms grouped together by types and probabilities. It gives the line context for every token, which can then be related to the correct wordclass or lemma using the interface shown in figure 1. This procedure supposes an excellent knowledge of the ancient language. Applied by a qualified scholar it resolves more than 90% of the morphological or lexical ambiguities. The remaining insecurities correspond mostly to philological or linguistic problems which can be resolved by an examination of the large context, a consideration of the textual habits or a consultation of the lexicographic and grammaticographic literature.

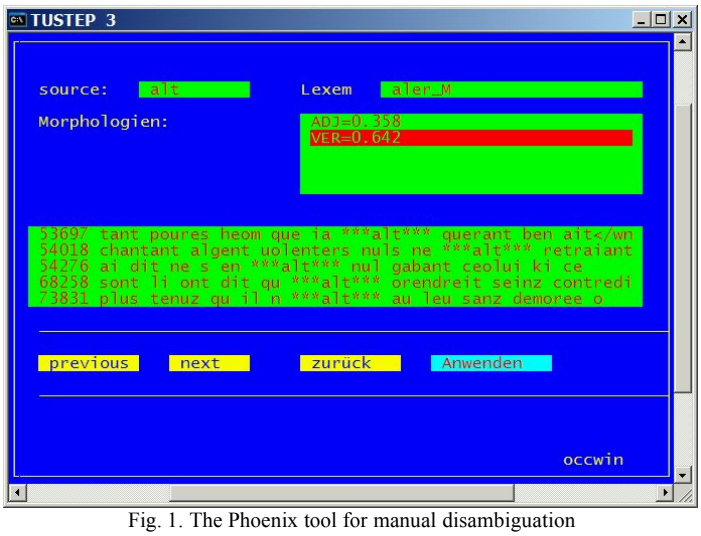

Fig. 1. The Phoenix tool for manual disambiguation

Depending of the corpus size, the semi-automatic disambiguation requires a large investment of time (about 50 types and about 500-700 tokens can be treated per hour). However, its results can help to refine the probabilistic procedure (e.g. by defining a threshold for low probabilities without any impact) and reduce the number of ambiguous cases in growing corpora.
Finally, the results of the lemmatization process might be considerably improved by defining the graphical equivalences that are extremely frequent in the ancient language (e.g. *y* = *i*, *ca-* = *ka*, double consonants or vowels = simple consonants or vowels, see Gleßgen 2003b). In a further application of

the principles of Phoenix, the unidentified forms could be assimilated to already lemmatized forms by means of the equivalences mentioned above.

2.6. Syntactic analysis

Two methods of automatic syntactic analysis have to be distinguished: Treebank parsers are trained on previously annotated texts and perform a purely probabilistic analysis. These parsers are normally fast and robust, but their precision depends heavily on the training corpus and the number of syntactic categories. Such a parser has been used for the annotation of the *Penn-Helsinki Parsed Corpus of Middle English* (Kroch/Taylor 2000). The second type of parsers use a grammar and a lexicon, and their analysis is based on syntactic rules. This approach presupposes the decision for a specific syntactic framework, and it requires even more lexical information than part of speech tagging, for example the valency of verbs. Some tests have already been made with the YAP parser (Schmid 2000) and a GB grammar which explicitly implements movement rules. The output file of YAP is plain text, but it can also be graphically visualized as shown in fig. 2, where the second sentence of the *Chanson de Roland* is analysed as a structure with inversion of the (empty) subject, after PP topicalisation, with the verb (*cunquist*) in the complementizer position (see M. Becker's contribution in this volume for a discussion of VSO structures in Old French).

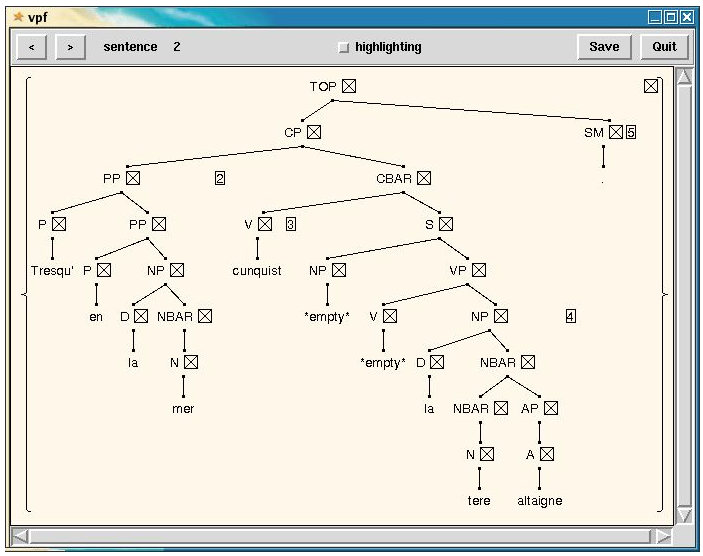

Fig. 2. YAP parser output

The disadvantage of grammar-based syntactic analysis is that it generates ambiguous structures which have to be disambiguated manually (for example the graphical tool shown here, by selecting one out of several trees) or probabilistically. However, resources and manpower needed for the syntactic annotation of a corpus of the size of the AC make it a long term project which is currently not in the reach of the SOFA project.

3. Conclusion and outlook

We have shown how various lexical and textual resources for Old French have been combined and treated with tools for automatic linguistic analysis in order to (1) enhance the quality of the resources, (2) extract lexical information of corpora, and (3) build specific tools for the automatic treatment of Old French. The SOFA project is still work in progress. At the next stages, the tools for part of speech annotation and lemmatization of Old French will be improved, and an annotated and reviewed version of the Amsterdam Corpus will be made available by 2006.

Bibliography

Blumenthal, Peter / Stein, Achim (eds.) 2002: *Electronic edition of the Altfranzösisches Wörterbuch von Tobler, Lommatzsch u.a.*. Stuttgart: Franz Steiner Verlag.

Dees, Anthonij 1987: *Atlas des formes linguistiques des textes littéraires de l'ancien français.* Tübingen: Niemeyer.

Gleßgen, Martin-Dietrich 2003a: L'élaboration philologique et l'étude lexicologique des 'Plus anciens Documents linguistiques de la France' à l'aide de l'informatique; in: Duval, Frédéric (ed.): *Actes du Xe colloque international sur le moyen français (12-14 juin 2000, Metz).* Paris: École des Chartes, 371-386.

Gleßgen, Martin-Dietrich 2003b: La lemmatisation de textes d'ancien français: méthodes et recherches; in:Kunstmann, Pierre et. al. (eds.): A*ncien et moyen français sur le Web: Enjeux méthodologiques et analyse du discours.* Ottawa: Les Éditions David, 273-284.

Kroch, Anthony / Taylor, Ann (eds.) 2000: *The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2).* Philadelphia: University of Pennsylvania.

Godefroy, Frédéric 1880: *Dictionnaire de l'ancienne langue française et tous ses dialectes*, Paris.

Huguet, Edmond 1925ff: *Dictionnaire de la langue française du seizième siècle.* Paris: Champion.

Kunstmann, Pierre et. al. (eds.): A*ncien et moyen français sur le Web: Enjeux méthodologiques et analyse du discours.* Ottawa: Les Éditions David.

Schmid, Helmut 1994: Probabilistic Part-of-Speech Tagging using Decision Trees; in:Sima'an, K. / Bod, R. / Krauwer, S. / Scha, R. (eds.): *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP'94), Manchester September 1994*. Manchester: UMIST.

Schmid, Helmut 2000: *YAP: Parsing and Disambiguation With Feature-Based Grammars.* Phd. Thesis Universität Stuttgart.

Stein, Achim / Schmid, Helmut 1995: Etiquetage morphologique de textes français avec un arbre de décisions. *traitement automatique des langues*, vol. 36, no. 1-2: Traitements probabilistes et corpus, 23-35.

Stein, Achim 2003: Étiquetage morphologique et lemmatisation de textes d'ancien français; in:Kunstmann, Pierre et. al. (eds.): A*ncien et moyen français sur le Web: Enjeux méthodologiques et analyse du discours.* Ottawa: Les Éditions David, 273-284.

Tobler, Adolf / Lommatzsch, Erhard 1925ss: *Altfranzösisches Wörterbuch.* Berlin u.a.: Weidmann.

WWW addresses

Laboratoire de Français Ancien (LFA, Ottawa):
http://www.uottawa.ca/academic/arts/lfa/

Tobler/Lommatzsch *Altfranzösisches Wörterbuch*:
http://www.uni-stuttgart.de/lingrom/stein/tl/

TreeTagger: Old French Parameter Files
http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html