CORPUS

Corpus

7 | 2008
Constitution et exploitation des corpus d'ancien et de moyen français

Syntactic Annotation of Old French Text Corpora

Achim Stein

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/1510
DOI : 10.4000/corpus.1510
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 10 novembre 2008
ISSN : 1638-9808

Référence électronique
Achim Stein, « Syntactic Annotation of Old French Text Corpora », *Corpus* [En ligne], 7 | 2008, mis en ligne le 12 novembre 2009, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/1510 ; DOI : https://doi.org/10.4000/corpus.1510

Syntactic Annotation of Old French Text Corpora

Achim STEIN
Institut für Linguistik / Romanistik (Stuttgart)[1]

1. Previous Work: Resources and Annotation

In this survey, we focus on Medieval French and English corpora to highlight the discrepancy between the resources in both languages. We are nevertheless aware of ongoing work, especially for Romance corpora (Latin Dependency Treebank (Bamman and Gregory 2006, Perseus Project[2], Medieval Portuguese[3], etc.).

1.1. Corpora for Medieval English

For medieval English, important corpus resources have been built in various projects. Today, they constitute the reference in the domain of syntactically annotated medieval corpora and have become an important base for theoretical research in diachronic syntax.

Two corpora, the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE, Taylor *et al.* 2003) and the Penn-Helsinki Parsed Corpus of Middle English (PPCME2, Kroch and Taylor 2000) have been annotated either at the University of Pennsylvania (UPenn, project directed by A. Kroch) or using the probabilistic parsing method developed there. They are publicly available (but the PPCME2 is not free of charge) and distributed with *CorpusSearch*, a search tool conceived for these (and only these) corpora. The probabilistic parsing method, although leading to a robust automatic tool, implies important cost (pre-tagging, manual correction) for building the

1 Heilbronner Str. 7, D-70174 Stuttgart, achim.stein@ling.uni-stuttgart.de

2 http://www.perseus.tufts.edu/

3 http://www.ime.usp.br/~tycho/

Corpus n°7 « Constitution et exploitation des corpus d'ancien et de moyen français » (2008), 157-171

training data, and for post-processing (manual correction) the output, at least if precision is an issue.

The syntactic model uses a « limited tree representation in the form of labelled parentheses ».[4] The structure in (1) shows an annotated sentence from the PPCME2.[5]

(1)
```
 I fond the day of the month in manere as I seide;
(0
(1 IP-MAT
    (2 NP-SBJ (3 PRO I))
    (4 VBD fond)
    (5 NP-OB1 (6 D the)
        (7 N day)
        (8 PP (9 P of)
           (10 NP (11 D the)
               (12 N month))))
    (13 PP (14 P in)
        (15 NP (16 N manere)
            (17 PP (18 P as)
             (19 CP-ADV (20 WNP-1 0)
                  (21 C 0)
                  (22 IP-SUB (23 NP-OB1 *T*-1)
                       (24 NP-SBJ (25 PRO I))
                       (26 VBD seide))))))
    (27 E_S ;))
(28 ID CMASTRO,669.C1.198))
```

1.2. Corpora for Medieval French

In the past few years, two independent projects have built the two largest text corpora for Medieval French, and made them available for research: the *Base de Français Médiéval* (BFM), initiated in 1989 and since then augmented and enhanced by C. Marchello-Nizia and her team (ENS-LSH Lyon, see Guillot & Marchello-Nizia & Lavrentiev 2007), and the *Nouveau Corpus d'Amsterdam* (NCA, University of Stuttgart, see Stein

4 http://www-users.york.ac.uk/~lang18/Documentation/syn-ppcme2-lite.htm

5 See http://www.ling.upenn.edu/mideng for further examples and documentation

et al. 2006). In section 2 of this article, we will discuss the annotation of examples taken from the NCA.

The *Nouveau Corpus d'Amsterdam* (NCA, Stein et al. 2006) is the first edition of the Old French data collected in the 1980s in the research project of Anthonij Dees for his *Atlas des formes linguistiques des textes littéraires de l'ancien français* (Dees 1987). In a joint project[6] with P. Kunstmann (Ottawa University) the 3,3 million word corpus has been edited in a POS-tagged, lemmatised and XML-formatted version (Stein *et al.* 2006). The 300 texts and text extracts were selected according to Dees's principles in order to cover the regional varieties of Old French (see Kunstmann & Stein (2007a) for a general description of the NCA and Stein (2007) for technical information on POS-tagging, lemmatisation and XML markup).

1.3. Corpora and Studies on Diachronic Syntax

A syntactic reference corpus for Medieval French is a desideratum not only for French corpus linguistics, but for comparative typology and diachronic syntax, where corpus-based approaches and the consideration of older periods have become increasingly important.

The publication of the corpora for Medieval English has had a striking impact on morphological and syntactic research, directly or indirectly connected to these corpora, e.g. Kroch & Taylor (1997), Pintzuk *et al*. (2000), Kroch (2001), Trips (2002), Pintzuk & Taylor (2006), Haeberli & Ingham, (2007), Trips (2007), Stein & Trips (2008). Comparable publications on French diachronic syntax were either lacking an empirical base entirely or based on a small number of texts and could therefore, although of high quality, not discuss important typological questions on an empirical base of comparable size; we only cite Roberts (1993), Marchello-Nizia (1996), Vance (1997), Esperling (2001), Kaiser (2002), Becker (2005) for many others.

For French and other Romance languages, J. Meisel's project « Mehrsprachigkeit als Ursache und Folge von

6 Alexander-von-Humboldt-Stiftung, Transcoop grant III-DEU1112357: « Computergestützte Analyse und Edierung alt- und mittelfranzösischer Texte ».

Sprachwandel: historische Syntax romanischer Sprachen »[7] has collected and analysed syntactic data for selected phenomena in the diachronic development of French and other Romance languages. Although the project has made an important contribution to this domain and part of the research was corpus-based, neither elaboration nor annotation of corpora were intended: a small number of texts was analysed syntactically, and the sentences were stored in a database in order to extract examples for specific syntactic properties (see Kaiser & Meisel 1991, Rinke 2003, Goldbach 2006).

1.4. Automatic Parsing

In the Canadian GTRC project *Les voies du français*[8], Kroch's probabilistic parsing methodology is currently being applied to Old French texts. Martineau *et al.* (2007) describe the intended results for French. Syntactic structures produced in this project will be similar to the Middle English example (1) cited in section 1.1.

Didier Bourigault has built *Syntex*, a partial dependency parser for Modern French. It analyses noun phrases and verb phrases. However, Bourigault *et al.* (2005) have shown that the choice of a syntactic representation does not imply the choice of a particular method of automatic analysis. They also show that annotated dependencies can easily be translated to constituent structures by conversion of dependency types into labelled edges, so that existing search tools like *TigerSearch* (Lezius 2002) can be used for corpus queries.

A similarity between our approach of manual pre-annotation and *Syntex* is that our syntactic annotation also builds on POS tagging (*Syntex* uses POS annotated texts tagged by the French TreeTagger, cf. Stein & Schmid 1995). Therefore *Syntex* is theoretically an interesting tool – at least for partial syntactic pre-annotation. However, the technology is not at our disposal, since it was partly developed with private funding

7 This project is part of the German SFB 538 « Mehrsprachigkeit », cf. http://www.uni-hamburg.de/fachbereiche-einrichtungen/sfb538/projekth1.html

8 GTRC project 412-2004-1002, University of Ottawa, speaker: France Martineau.

(société Synomia) and is not freely available for research purposes.

To sum up, neither automatic parsing method meets our goal of producing publicly available and reusable resources and tools: A. Kroch's probabilistic parser is trained and can be used *only* within the GTRC project, and *Syntex* is unavailable for research for commercial reasons.

2. Syntactic Annotation of Old French Corpora

2.1. Principles for Syntactic Annotation

With respect to the grammar model used for annotation, our proposal is in line with most of the treebank approaches: we avoid decisions in certain cases of ambiguity and adopt as few theoretical assumptions as possible, with a preference for flat syntactical structures.

Our grammar model is situated between the two extremes applied in existing treebanks: constituency-style annotation, as in the Old and Middle English Corpora, and dependency-style annotation, as in Prague Dependency Treebanks (Hajicová 2002): on the one hand, in our annotation, certain phrases will be either automatically chunked or manually annotated as constituents (e.g. NP, AP, PP). On the other hand, grammatical functions will not be defined by constituency, but annotated directly by named functions or by dependencies. A similar hybrid annotation is used for the *Paris 7 French Treebank* (Abeillé 2003) and explained in more detail in Abeillé & Barrier (2004).

The short sentence in (2) will exemplify some of these principles:

(2) tresqu en la mer cunquist la tere altaigne
'as far as to the sea (he) conquered the high land'

In our analysis of (2),

a) noun phrases (here: the direct object *la tere altaigne*) are not annotated as a DP, but an NP, since functional categories are not heads in our annotation;

b) this object NP it is not embedded in a VP, but depends on (or is governed by[9]) *cunquist*;
c) the empty subject category is not annotated: pro drop sentences will be marked by attributes at sentence level (since defining the empty subject position by annotating an empty category would be impossible without adhering to disputable theoretical assumptions);
d) the adverbial phrase *tresqu en la mer* is not adjoined to (governed by) *cunquist*, but depends on the phrase node.

2.2. Automatic Pre-annotation

Compared to modern language corpora, the Old French resources are of limited size. Still, in its current format, the NCA has about 300.000 <s>-boundaries corresponding to either sentences in prose texts or verses. The BFM is of comparable size and expected to grow in the future. The time for manual annotation per sentence depends on many parameters (e.g. depth of annotation and performance of the tool for manual annotation (TMA)) and is therefore hard to estimate. But it is quite obvious that partial automatic annotation is necessary to avoid the project continuing for decades. Annotation practice will show later to what extent such pre-chunking can be helpful for manual annotation. The tool *NotaBene* (Mazziotta 2006) will have the functionality to deal with pre-annotated structures, and in some cases, e.g. complex phrases, the pre-annotation might speed up the manual annotation process, whereas in others, the annotators might simply delete the chunks altogether.

In this contribution we are not going to discuss the different methods for syntactic pre-annotation[10]; rather, we will focus on concrete syntactic questions. That is why we chose to

9 In the terminology used here, « A governs B » and « B depends on A » are equivalent and express the same dependency relation between two elements A and B.

10 If probabilistic parsers are ruled out for methodological reasons, XSLT (« Extensible Style Sheet Transformations ») would certainly be an adequate choice which has already been adopted for similar projects dealing with dependency annotation (e.g. LASSY « Large Scale Syntactic Annotation of written Dutch »).

make first tests using our own tool, a Perl tool based on the TWIC (Tagged Words in Context) search tool originally designed for querying part-of-speech-tagged and XML formatted text corpora.[11] TWIC's « chunker » function can annotate chunks based on POS tag definitions. A first version handles frequent types of AP, NP, and PP phrases, as well as some verb and determiner clusters (i.e. unstructured sequences of verb and determiner types). It produces chunks like the ones shown in (6).

The chunker function allows the user to assign syntactic markup codes to these queries by formulating rules like the one in (3), which states that markup for a noun phrase (NP) will be introduced for each sequence of determiner, adjective and noun. The options on the left side of the rule tell the chunker to include the name of the rule (option « r ») and the chunked string (« s ») in the markup and to apply the rule only if the constituents agree in number, gender and case (« a »).

(3) `NP,det-a-n,rsa -> pos=DET pos=ADJ pos=NOM`

TWIC does not compile a grammar from these rules, but simply processes them one by one, in the order in which they are specified. A rule can therefore build on the syntactic markup inserted by previous rules, as in (4), where the attribute « xml » on the right side refers to a previously inserted category, but the markup resulting from subsequent rules is of course not available at processing time.

(4) `NP,det-a-np -> pos=DET pos=ADJ xml=NP`

Other limitations of this method come from the rule input: the chunker relies on the part of speech annotation, which means that the rules have to account for errors and particuliarities like for example the underspecified category « PROCON » (originally « 600 » in A. Dees's annotation scheme) which is used for pronouns, conjunctions and some adverbs.

11 TWIC is the query engine used for online access to version 2 of the *Nouveau Corpus d'Amsterdam*.

In the current experimental phase, we apply about thirty rules for complex verb forms (VCOMP), adjective phrases (AP), noun phrases (NP) and prepositional phrases (PP). Our goal is to speed up the manual tagging of the corpus by pre-annotating frequent chunks in a reliable way, favouring precision over recall since manual corrections would be very time-consuming. We therefore abandoned rules which produced too many errors, like « NP, det-n-a », because many adjectives in postposition do not actually depend on the preceding noun:

(5) sun braz teneit *[le chef enclin]

The following table lists the rules ranked by their frequency of application in a sample corpus (*La chanson de Roland*: about 30.000 words, 2558 sentences).

```
1506 NP, det-n -> pos=DET pos=NOM
 794 PP, pre-n -> pos=PRE pos=NOM
 560 PP, pre-np -> pos=PRE xml=NP
 301 NP, nodet-a-n -> pos=ADJ pos=NOM
 283 VCOMP, aux-pper -> pos=VER:VFLEKT pos=VER:pper
 160 VCOMP, aux-infi -> pos=VER:VFLEKT pos=VER:infi
 153 VCOMP, se-v -> se?&pos=PROCON pos=VER
 146 AP, adv-adj -> pos=ADV pos=ADJ
 126 NP, det-a-n -> pos=DET pos=ADJ pos=NOM
 117 PP, pre-propers -> pos=PRE pos=PRO:pers
  92 NP, np-npr -> xml=NP,rule=det-n pos=NPR
  81 DP, ind_det -> pos=DET:ind pos=DET
....
```

Table 1: Rules ranked by frequency of application

The 30 rules of our experimental « grammar » provide markup for more than 6.000 annotated phrases, which is a good start for the manual annotation of dependencies at a higher level, since the most common noun and prepositional phrases will have been annotated, and the annotator will be able to focus on the more complex tasks.

Example (6) shows the output of the chunker for sentence (2) discussed in section 2.1. and its preceding context. The XML elements inserted by the chunker have been prefixed by « CHUNK- » to distinguish them from existing elements. Some default attributes (« rule » to indicate the chunking rule) have been eliminated here for better readability. The chunker has inserted the attributes which allow to build the dependency structure: « id » is a unique identifier for each phrase, and

« gov » indicates the id of the governing phrase and must be filled by a phrase id (the default being the top node, i.e. the sentence id). The annotator corrects the « gov » values where necessary and thus builds the dependency structure for the whole sentence, either manually or, ideally, assisted by the annotation tool. The chunker also inserts the attribute « func », for the grammatical function of the phrase, and can suggest a default value based on the phrase type (« mod(ifier) » for AP and PP) or on the feature information (« suj(et) » for subject case, « obj(et) » for oblique case).

(6) pre-annotation of: set anz tuz pleins ad estét en espaigne / tresqu en la mer cunquist la tere altaigne

```
<s id="s2" line="2">
<CHUNK-NP id="s2.np1" gov="s2" func="obj">
  <word pos="ADJ">set</word>
  <word pos="NOM">anz</word>
</CHUNK-NP>
<CHUNK-AP id="s2.ap1" gov="s2.np1" func="mod">
  <word pos="PRO">tuz</word>
  <word pos="ADJ">pleins</word>
</CHUNK-AP>
<CHUNK-VCOMP id="s2.v1" gov="s2">
  <word pos="VER">ad</word>
  <word pos="VER">estét</word>
</CHUNK-VCOMP>
<CHUNK-PP id="s2.v1" gov="s2.v1" func="mod">
  <word pos="PRE">en</word>
  <word pos="NPR">espaigne</word>
</CHUNK-PP>
</s>
<s id="s3" line="3">
<word pos="ADV">tresqu</word>
<CHUNK-PP id="s3.pp1" gov="s3" func="mod">
  <word pos="PRE">en</word>
  <CHUNK-NP id="s3.np1" gov="s3.pp1" func="obj">
    <word pos="DET">la</word>
    <word pos="NOM">mer</word>
  </CHUNK-NP>
</CHUNK-PP>
<word pos="VER">cunquist</word>
<CHUNK-NP id="s3.np2" gov="s3" func="obj">
  <word pos="DET">la</word>
  <word pos="NOM">tere</word>
</CHUNK-NP>
<word pos="ADJ">altaigne</word>
</s>
```

2.3. Specification of the Annotation Tool

We have shown that automatic pre-annotation can complement manual annotation, and that the extent to which this technique will be applied depends on several factors (depth of annotation, precision of the chunker rules, skills of human annotators, etc.).

A tool for manual annotation (TMA) should therefore be as flexible as possible. We therefore conclude our reflexions on manual annotation by stating some of the requirements which a TMA appropriate for a large-scale annotation project should meet[12]:

1 System
   a) The TMA must be fast with regard to the user interaction (response to input, display of modifications, key and mouse buffering, etc.). The manual annotator should never have to wait for a process to finish.
   b) As system-independent as possible (run on Linux, Mac and Windows machines and be integrated as well as possible into their surface with regard to standard keystrokes, etc.).

2 User Interface
   a) Keyboard input is usually safer and faster than mouse input. The TMA should permit both types of input for the selection of text and the assignment of categories.
   b) The possibility of defining shortcuts for keyboard and mouse actions is essential to speed up the manual annotation.
   c) The user should be able to simplify the tree representation (fold and unfold substructures, e.g. embedded phrases) so that complex sentences will be easier to read.
   d) The interface should be in English or provide methods for localisation.

3 Syntactic Annotation
   a) The TMA should accept corpora with different degrees of previous syntactic annotation (manual or automatic) and thus allow increases in annotation depth.
   b) Deletion and correction of previous annotation must be allowed for. Optionally, the TMA could keep track of

12 Our goal is not to give a complete description of the functionality of the TMA. Note that the tools *KhEdit* and *Notabene* already meet some of these requirements (Mazziotta 2006).

deletions and corrections (by creating comments or writing them to a log file).
   c) Dependencies between words and categories must be assignable as stated above. Default assignments (e.g. to the sentence node) should be possible.
   d) The names of the inserted information (categories, functions, etc.) and their assignment to XML codes (elements, attributes, values) must be configurable.
   e) Discontinuous elements (e.g. auxiliary and main verb separated by an NP) must be selectable.
   f) The user should be able to insert comments at category level (within an XML element, e.g. to comment on a particular choice) and freely in the text (XML comment, e.g. with regard to a particular structure).

4 Input/Output
   a) Support for standard character encoding (ISO Latin and Unicode).
   b) Regular backups should be configurable.
   c) The output of discontinuous elements should be configurable: either linear output (in word order, the annotation indicating the dependencies) or clustered (the annotation indicating the original position).

5 Options (useful, not mandatory):
   a) Save preferences for different users or different corpora.
   b) Provide configurable input/output formats or interfaces with scripts for input/output (e.g. export for query tools like *TigerSearch*).
   c) Control input for free text (XML #CDATA) to avoid typing errors.
   d) Definition of the annotation depth, so that the tool displays if the current annotation is complete or not.

3. Conclusion

In contrast to other syntactic annotation projects, our goal is to annotate the *Nouveau Corpus d'Amsterdam* (NCA) manually, assisted by automatic pre-annotation (chunking). We have

presented various projects for medieval corpora and discussed some examples for the interaction between pre-annotation, manual annotation, and the respective tools for these tasks.

Further development of the project will depend on public funding and on cooperation with similar annotation initiatives for medieval French corpora, most of all the *Base de Français Médiéval* (BFM, Lyon). A joint German-French research proposal has therefore been submitted to ensure that NCA and BFM will be annotated according to the same principles, so that the future users will get maximum benefit from this research programme.

References

Abeillé A. (éd.) (2003). *Treebanks : building and using parsed corpora*. Dordrecht : Kluwer.

Abeillé A. & Barrier N. (2004). « Enriching a French Treebank ». In 4th international conference on language resources and evaluation LREC. Lisbon.

Bamman D. & Crane G. (2006). « The Design and Use of a Latin Dependency Treebank ». In *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories* (TLT, Prague 2006), 67-78.

Becker M. (2005). « Le 'Corpus d'Amsterdam' face à une vieille question : l'ancien français est-il une langue V2 ? ». In J. Kabatek, C. Pusch & W. Raible (eds) *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*. Tübingen : Narr, 345-358.

Bourigault D., Fabre C., Frérot C., Jacques M.-P. & Ozdowska S. (2005). « Syntex, analyseur syntaxique de corpus ». In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*. Dourdan.

Dees A. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français*. Avec le concours de M. Dekker, O. Huber et K. van Reenen-Stein. Tübingen : Niemeyer.

Esperling P. (2001). *Untersuchungen zur Syntax (Wortstellung) im Mittelfranzösischen des 15. Jahrhunderts*. Frankfurt a.M. : Lang.

Goldbach M. (2006). « Kontrastiver Vergleich der syntaktischen Verteilungen der starken und schwachen Objektpronomen im Alt- und Mittelfranzösischen und im Altitalienischen », *Zeitschrift für Romanische Philologie* 122 : 380-416.

Guillot C., Marchello-Nizia C. & Lavrentiev A. (2007). « La base de français médiéval (BFM) : états et perspectives ». In P. Kunstmann & A. Stein (eds) *Le Nouveau Corpus d'Amsterdam*. Stuttgart : Steiner.

Haeberli E. & Ingham R. (2007). « The position of negation and adverbs in Early Middle English », *Lingua* 117 : 1-25.

Hajicová E. (2002). « Theoretical description of language as a basis of corpus annotation. The case of the Prague dependency Treebank ». In E. Hajicová *et al.* (eds) *Travaux du Cercle Linguistique de Prague* n.s. / *Prague Linguistic Circle Papers*. Amsterdam : Benjamins, volume 4 : 111-127.

Kaiser G. (2002). *Verbstellung und Verbstellungswandel in den romanischen Sprachen*. Tübingen : Niemeyer.

Kaiser G. & Meisel J. (1991). « Subjekt und Null-Subjekt im Französischen ». In S. Olsen & G. Fanselow (eds) *DET, COMP und INFL. Zur Syntax funktionaler Kategorien und grammatischer Funktionen*. Tübingen : Niemeyer, 110-136.

Kroch A. (2001). « Syntactic change ». In M. Baltin & C. Collins (eds) *The Handbook of Contemporary Syntactic Theory*, chapter 22. Oxford : Blackwell.

Kroch A. and Taylor A. (1997). « Verb movement in Old and Middle English: Dialect variation and language contact ». In A. V Kemenade, & N. Vincent (eds) *Parameters of Morphosyntactic Change*. Cambridge : Cambridge University Press, 297–325.

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

Kunstmann P. & Stein A. (2007a). « Le Nouveau Corpus d'Amsterdam ». In P. Kunstmann & A. Stein (eds) *Le Nouveau Corpus d'Amsterdam*. Actes de l'atelier de Lauterbad, 23-26 février 2006. Stuttgart : Steiner, 9-27.

Kunstmann P. & Stein A., (eds) (2007b). *Le Nouveau Corpus d'Amsterdam.* Actes de l'atelier de Lauterbad, 23-26 février 2006. Beihefte zur Zeitschrift für französische Sprache und Literatur 34. Stuttgart : Steiner.

Lezius W. (2002). *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora (German)*. University of Stuttgart Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), vol. 8, n° 4. Stuttgart : Institut für Maschinelle Sprachverarbeitung (IMS).

Marchello-Nizia C. (1996). « Les verbes supports en diachronie. Le cas du français », *Langages* 121 : 91-98.

Martineau F., Diaconescu C. & Hirschbühler P. (2007). « Le corpus 'voies du français' : de l'élaboration à l'annotation ». In P. Kunstmann & A. Stein (eds) *Le Nouveau Corpus d'Amsterdam*. Actes de l'atelier de Lauterbad, 23-26 février 2006. Stuttgart : Steiner, 121-142.

Mazziotta N. (2006). « Objectifs, méthodes et outils du projet Khartês ». In A. Mardale, M. Soufflard & G. Vecherand (eds) 9ème Atelier des Doctorants en Linguistique. 17 & 18 octobre 2005. Paris : Université de Paris 7, 188-195.

Pintzuk S. & Taylor A. (2006). « The loss of OV order in the history of English ». In A. van Kemenade & B. Los (eds) *Blackwell Handbook of the History of English*, chapter 11. Oxford : Blackwell.

Pintzuk S., Tsoulas G. & Warner A. (eds) (2000). *Diachronic Syntax. Models and Mechanisms*. Oxford : Oxford University Press.

Rinke E. (2003). « On the licensing of null subjects in Old French ». In U. Junghanns & L. Szucsic (eds) *Syntactic*

Structures and Morphological Information. Berlin / New York : Mouton de Gruyter, 217-249.

Roberts I. (1993). *Verbs and diachronic syntax. A comparative history of English and French*. Dordrecht : Kluwer.

Stein, A. (2007) « Resources and Tools for Old French text corpora ». In Y. Kawaguchi *et al.* (eds) *Corpus-Based Perspectives in Linguistics*. Amsterdam / Philadelphia : Benjamins, 217-229 (= Usage Based Linguistic Informatics 6).

Stein A. *et al*. (eds) (2006). *Nouveau Corpus d'Amsterdam*. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen, Stuttgart : Institut für Linguistik / Romanistik.

Stein A. & Schmid H. (1995). « Étiquetage morphologique de textes français avec un arbre de décisions », *Traitement automatique des langues* 36, 1-2 : 23-35.

Stein A. & Trips C. (2008). « Was Old French '-able' borrowable ? A diachronic study of word-formation processes due to language contact ». In M. Dossena, R. Dury & M. Gotti (eds) *English Historical Linguistics*. Volume 2: *Lexical and Semantic Change*. Amsterdam / Philadelphia : Benjamins.

Taylor A. *et al*. (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)*. York : University of York, Heslington.

Trips, C. (2002). *From OV to VO in Early Middle English*. Amsterdam / Philadelphia : Benjamins.

Trips C. (2007). Lexical semantics and diachronic morphology: the development of the derivational suffixes '-hood', '-dom' and '-ship' in the history of English, Universität Stuttgart: habilitation thesis.

Vance B. (1997). *Syntactic Change in Medieval French. Verb-second and null subjects*. Dordrecht : Kluwer.