

Conference on New Methods in Historical Corpora
University of Manchester, 29-30 April 2011

Syntactic annotation of medieval texts. Theoretical and practical issues

Achim Stein and Sophie Prévost

CNRS Lattice–UMR 8094/ENS Paris, France
Institut für Linguistik/Romanistik, Universität Stuttgart, Germany

April 29, 2011

SRCMF: organisation and funding

- ▶ **Title:** *Syntactic Reference Corpus of Medieval French*
- ▶ **Funding:** Agence nationale de la recherche (ANR) and Deutsche Forschungsgemeinschaft (DFG), 1.3.2009-29.2.2012
- ▶ **Staff:** principal investigators**, researchers* and cooperators
 - CNRS Lattice Paris (F):** Sophie Prévost**, Julie Glikman*
 - ENS Lyon (F):** Céline Guillot, Serge Heiden, Alexei Lavrentiev*,
Christiane Marchello-Nizia (émérite), Tom Rainsford*
 - ILR Universität Stuttgart (D):** Achim Stein**, Beatrice-Barbara
Bischof*, Nicolas Mazziotta*

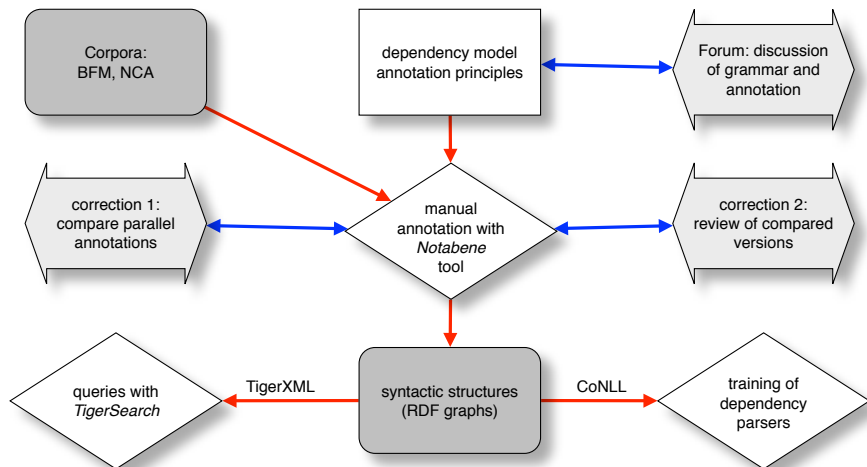
Corpus resources

- ▶ *Base de Français Médiéval* (BFM)
 - ▶ see Guillot et al. (2007) and <http://bfm.ens-lyon.fr/>
 - ▶ originally compiled by C. Marchello-Nizia (ENS Lyon)
 - ▶ 80 complete Old and Middle French texts (> 3 million words)
 - ▶ annotated with Cattex part of speech tagset (Dupuis et al., 2000):
http://bfm.ens-lyon.fr/article.php3?id_article=176
- ▶ *Nouveau Corpus d'Amsterdam* (NCA)
 - ▶ see Stein et al. (2006) and <http://www.uni-stuttgart.de/lingrom/stein/corpus/>
 - ▶ originally compiled by A. Dees (Amsterdam)
 - ▶ edited by Pierre Kunstmann (Ottawa), Achim Stein (Stuttgart) and Martin-Dietrich Glessgen (Zurich, bibliography).
 - ▶ 300 samples of OldFrench texts (> 3 million words)
 - ▶ part-of-speech (POS) annotated, lemmatised, Dees/ILR tagset

Goals

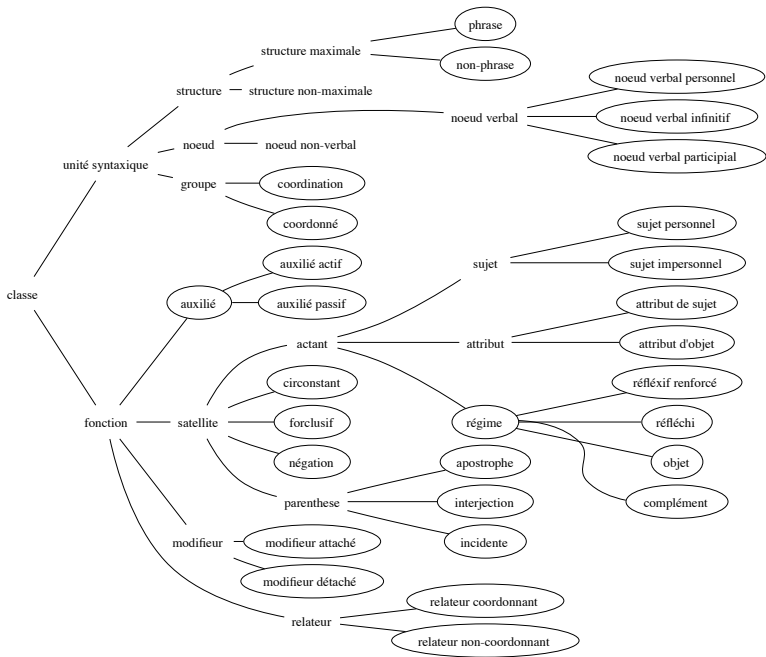
- ▶ **Manual syntactic annotation** of a subset of the BFM/NCA corpus
 - ▶ Development of a dependency-based grammar model
 - ▶ Bootstrapping of POS annotation from the syntactic annotation
 - ▶ correct the existing annotations
 - ▶ create a common POS annotation standard for both corpora (i.e. apply the BFM-Cattex tagset to the NCA)
- ▶ Construction of a **reference corpus** for syntactic research and training of tools (parsers).
- ▶ **Availability:**
 - ▶ Resources produced in SRCMF will be made available under a Creative Commons (CC) international licence, plus specific (national) licences if necessary.
 - ▶ Software produced in SRCMF is freely available.
Notabene annotation tool: <http://sourceforge.net/projects/notabene/>

SRCMF annotation workflow



The SRCMF grammar model

- ▶ Following the specifications of the *NotaBene* annotation tool (Mazziotta, 2010), SRCMF uses a class hierarchy for
 - ▶ structures: sentence, verb node, group (coordination)
 - ▶ functions: subject, attribute, modifier etc.
- (see the attached figure for an overview of the hierarchy.)
- ▶ A dependency relation is expressed by the triplet:
mother node – child node – dependency relation
 - ▶ for a similar approach see the *Turin University Treebank* (TUT; Bosco, 2004)



Heads and functional elements

Top nodes of structures

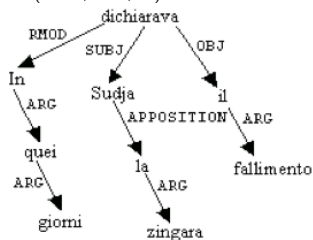
All the structures are headed by the lexical head (verb, noun, adjective, adverb). Lexical heads are preferred over functional heads as top nodes of a structure.

- ▶ Main clauses are headed by the inflected verb (or the first inflected element of the verb complex).
 - ▶ The matrix verb immediately dominates the verb of the subordinate clause
 - ▶ The functional category (e.g. the conjunction) depends on the verb
- ▶ Prepositional phrases are headed by the noun, the preposition depends on the noun.

Heads and functional elements

- ▶ In this respect, the SRCMF grammar differs from the *TUT* model, where functional categories nodes govern lexical heads:

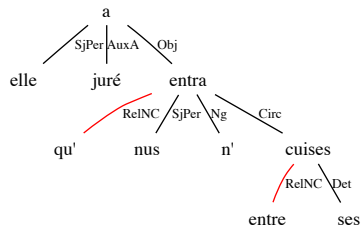
TUT (Bosco, 2004, 89):



In these days Sudja the gypsy declared the failure

RMOD = restrictive modifier

SRCMF:



She has sworn that nobody not entered between her thighs

Empty nodes and traces

Duplicates

A **duplicate** is the **copy of a node**. This allows to attach a second relation to the duplicate.

- ▶ Duplicates are used in relative clauses, interrogatives and contracted forms.
 - ▶ In (1), the relative pronoun *qui* is a non-coordinating relator (**ReINC**), its duplicate is a subject (**SjPer**)
 - ▶ In (2), the contracted form *nes* (*ne+les*) is a negation (**Ng**), its duplicate is an object (**Obj**).

(1) Souffrance si est semblable a esmeraude **qui** toz jorz est vert. (Graal v.17-18)
Sufferance such is like an emerald which all day is green.

(2) sovent dit / Qu'or veut morir s'il **nes** ocit. (TristBérM4 v.1985-1986)
*often says / that now wants die if he **not+them** kills*

NotaBene: features

NotaBene (Mazziotta, 2010) is a tool for manual syntactic annotation. It allows the user to

- ▶ create and modify the syntactic annotation using a graphical interface which allows the user to manipulate a tree structure;
- ▶ add free comments to any node of the structure, search and list them;
- ▶ use scripts for semi-automatic correction;
- ▶ create text-specific or user-specific annotations, which makes it easy to modify the names or labels of categories;
- ▶ compare two annotated versions of the same text and allow for parallel modification (see figure).

- └ Dependency annotation
- └ Annotation tools and procedures

NotaBene annotation tool: compare mode

The screenshot displays the NotaBene application window titled "Achim - 2/beroul 01/17 - NotaBene". The interface is divided into several panels:

- Text view:** Shows the original text with line numbers 39 to 41: "Et Dex l si ne m'en croit il pas .", "Je puis dire : de haut si bas !", "Sire , mot dist voir Salemon :".
- Tree view (Left):** Displays a syntactic tree for the text. The root node is "n" (srcmf:hierarchy), which branches into "NgPrt" (pas), "Snt" (je), "VFin" (puis), "AuxA" (dire), "Obj" (Adv), "ModA" (de), "Adv" (haut), "ModA" (si), and "Snt" (bas). The "Snt" node is highlighted with a blue circle.
- Tree view (Right):** Shows a second syntactic tree for comparison. It has a similar structure but with different node labels and highlights. The root is "n" (srcmf:hierarchy), branching into "Circ" (en), "VFin" (croit), "SjPer" (il), "NgPrt" (pas), "Snt" (je), "VFin" (puis), "AuxA" (dire), "Obj" (Adv), "ModA" (de), "RelNC" (haut), "Adv" (si), and "Adv" (bas). The "Snt" node is highlighted with a blue circle.
- Control panel:** Located on the right, it includes a "Terminology" section with a tree view showing a hierarchy of classes. The selected class is "srcmf:hierarchy", which includes sub-classes like "srcmf:Fonction", "srcmf:Auxile", "srcmf:Actant", etc.
- Search and Comparison:** A search bar at the top of the tree views contains "srcmf:hie". A "Compared" section shows "n" and "srcmf:hierarchy".
- Status Bar:** At the bottom, it indicates "Differences: 67" and "10" (likely the number of differences shown).

NotaBene: representation

- ▶ *NotaBene* uses graphs in XML/RDF format (resource description format; Bechhofer et al., 2004) for the internal representation of the annotation.
- ▶ *NotaBene* currently exports to the following formats:
 1. dot/GraphViz: for graph images
 2. TigerXML:
 - ▶ used by the *TigerSearch* query software (IMS, Stuttgart; Lezius, 2002)
 - ▶ allows us to distribute SRCMF corpora in a comfortable query format
 3. CoNLL:
 - ▶ defined by the Conference on *Computational Natural Language Learning* (e.g. CoNLL 2009 shared task)
 - ▶ the standard format for dependency parsers

- └ Dependency annotation
- └ Annotation tools and procedures

Dependency structures in TigerXML

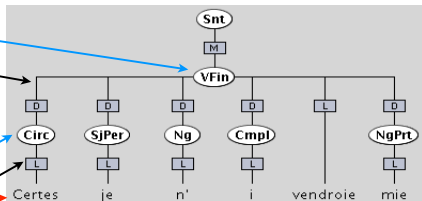
```

<?xml id='beroul_pb:1_lb:19_1263221020.72'>
  <graph root='_1263221020.72'>
    <terminals>
      <t word="Certes" id="w26_00095" pos="ADV" lemma="certes"/>
      <t word="je" id="w26_00097" pos="PRO_pers" lemma="je"/>
      <t word="n'" id="w26_00098" pos="PRO_clit" lemma="ne"/>
      <t word="i" id="w26_00099" pos="PRO_clit" lemma="y"/>
      <t word="vendroie" id="w26_00100" pos="VER" lemma="venir"/>
      <t word="mie" id="w26_00101" pos="ADV" lemma="mie"/>
    </terminals>
    <nonterminals>
      <nt id="_452409.15" cat="Ng">
        <edge label='L' idref='w26_00098' />
      </nt>
      <nt id="_221023.93" cat="VFin">
        <edge label='D' idref='_452410.38' />
        <edge label='D' idref='_452418.9' />
        <edge label='D' idref='_452406.05' />
        <edge label='D' idref='_452407.65' />
        <edge label='D' idref='_452409.15' />
        <edge label='L' idref='w26_00100' />
      </nt>
      <nt id="_452410.38" cat="NgPrt">
        <edge label='L' idref='w26_00101' />
      </nt>
      <nt id="_452418.9" cat="Cmpl">
        <edge label='L' idref='w26_00099' />
      </nt>
      <nt id="_452406.05" cat="Circ">
        <edge label='L' idref='w26_00095' />
      </nt>
      <nt id="_221020.72" cat="Snt">
        <edge label='M' idref='_221023.93' />
      </nt>
      <nt id="_452407.65" cat="SjPer">
        <edge label='L' idref='w26_00097' />
      </nt>
    </nonterminals>
  </graph>
</?xml>

```

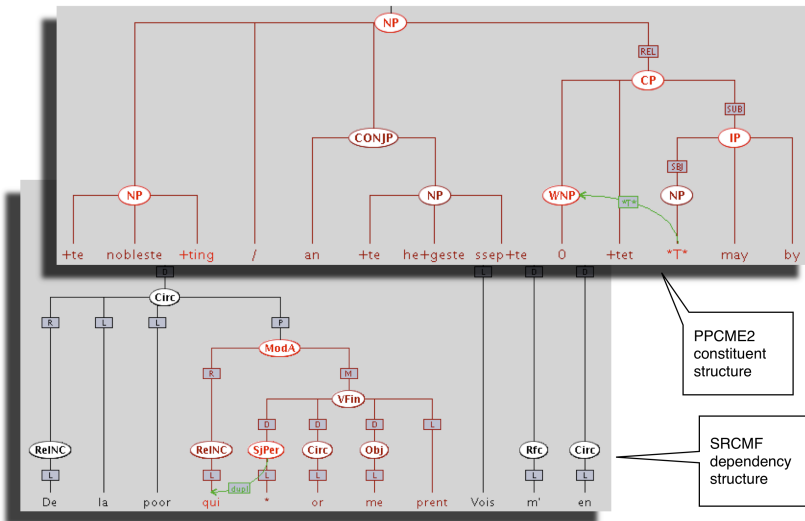
terminale Knoten

nicht terminale Knoten



- └ Dependency annotation
- └ Annotation tools and procedures

Dependency und constituency (in TigerSearch)



Perspectives

- ▶ Use our manual annotation as gold-standard training corpus for dependency parsers. Tests have been made with *mate-tools* (Bohnet, 2010; Björkelund et al., 2010).
- ▶ In the project *CoSToMeF* (submitted to ANR/DFG: 2012-2015):
 - ▶ integrate dependency parsing and manual annotation (*NotaBene*) into a graphical corpus environment *TXM*
 - ▶ <http://textometrie.ens-lyon.fr/> (Heiden, 2010)
 - ▶ extend lexical statistic functions to syntactic statistics
 - ▶ provide an easy-to-use web interface

- Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., F., P.-S. P., and Andrea Stein, L. (2004). *OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004*. W3C, <http://www.w3.org/TR/owl-ref/>.
- Björkelund, A., Bohnet, B., Hafdehl, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, COLING '10*, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bohnet, B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China. Coling 2010 Organizing Committee.
- Bosco, C. (2004). *A Grammatical Relation System for Treebank Annotation*. PhD Thesis, Università degli Studi di Torino.
- Dupuis, F., Heiden, S., and Prévost, S. (2000). Catégorisation d'un corpus hétérogène de français médiéval. Actes 5e journées internationales d'Analyse Statistiques des Données Textuelles (JADT'2000), Lausanne. JADT.
- Guillot, C., Marchello-Nizia, C., and Lavrentiev, A. (2007). La base de français médiéval (bfm) : états et perspectives. In Kunstmann, P. and Stein, A., editors, *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*. Steiner, Stuttgart.
- Heiden, S. (2010). The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24). 4-7 November 2010, Sendai*.
- Lezius, W. (2002). *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora (German)*. University of Stuttgart Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), vol. 8, no. 4. Institut für Maschinelle Sprachverarbeitung (IMS), Stuttgart.
- Mazziotta, N. (2010). Building the 'syntactic reference corpus of medieval french' using notabene rdf annotation tool. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*.
- Stein, A. et al., editors (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*. Institut für Linguistik/Romanistik, Stuttgart.

THANK YOU!