Syntactic Reference Corpus of Medieval French (SRCMF)

June 8, 2016

Sophie Prévost (CNRS Lattice, Paris) / Achim Stein (ILR, Universität Stuttgart)

1 SRCMF

1.1 Reference to the corpus

- Please cite the corpus as follows:
  Prévost, Sophie; Stein, Achim. 2013. Syntactic Reference Corpus of Medieval French (SRCMF) [version number]. ENS de Lyon/ILR Stuttgart. BibTeX entry
- The SRCMF is registered as International Standard Language Resource with the ISLRN 899-492-963-833-3

1.2 About the project

- **Funding:** *Syntactic Reference Corpus of Medieval French (SRCMF)* was financed by the 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*
- **Description:** The SRCMF is the first dependency treebank for Old French. It consists of syntactically annotated parts of two text corpora of Medieval French:
  - *Base de Français Médiéval* (BFM), see Guillot et al. (2007) and http://bfm.ens-lyon.fr/
  - *Nouveau Corpus d'Amsterdam* (NCA), see Stein et al. (2006); Kunstmann & Stein (2007) and http://www.uni-stuttgart.de/lingrom/stein/corpus/

A treebank is a text corpus which includes the annotation of a syntactic structure for each sentence. For SRCMF, a dependency grammar was used. Fifteen texts covering the Old French period from 842 to the end of the 13th century and containing about 251 000 words were annotated manually and published along with the tools and documentation on this website. The work group delivered the following results and resources:

- a dependency grammar model for the annotation of these corpora;
- a detailed documentation of the syntactic categories of the grammar model, as well as annotation guidelines;
- the tool *NotaBene* for manual syntactic annotation and comparison of annotated corpora;
- a double-checked, high quality morphological and syntactic annotation
- interfaces to export the annotation to TigerXML (allowing to use the *TigerSearch* tool) and to the CoNLL shared task format (allowing to train dependency parsers).

1.3 Quick start

If you already know TigerSearch (or if you don't want to read instructions):

1. Download and install TigerSearch (see section 4.2 below).
2. Download the binary files of a corpus text and install them (see section 2.1 below).

You may also wish to use this detailed tutorial:

- The SRCMF tutorial (PDF)

2 Access

2.1 Online access

(in preparation)

Online queries via a TXM web interface for SRCMF will be provided by the database Base de français médiéval (BFM). This TXM interface will allow for TigerSearch queries online.

2.2 Download

In this section, you can download the SRCMF files in various formats.

Note that due to copyright restrictions enforced by the publisher (Droz), texts which can't be licensed for download will not contain word forms. However, syntactic structures will be available for *all* the texts, and the lack of word forms will be compensated by the addition of other types of information (morphological markup, lemma) and references to the original text.

This table does not contain detailed information about the corpora. Please refer to the bibliographies of BFM and NCA (see section 1.2).

- **Title/Date**: Short title and approximate date. For details consult the BFM bibliography or the date given in the header of the XML-TEI edition.
- **Words**: Rounded figures indicate that the text is an excerpt of the complete original text (e.g. *Roman de la rose*).
- **TigerBinary** files: ready to use with TigerSearch. Unpack the zipped files, move the resulting directory in the directory *TIGERCorpora* of your TigerSearch installation. The corpus will appear if you (re)start TigerSearch or if you refresh the corpus tree if TigerSearch is already running.
- **TigerXML**: install TigerSearch, unpack the zipped XML files, use TigerRegistry (delivered with TigerSearch) to register them in TigerSearch. Use TigerSearch for queries (see the TigerSearch manual for more information).
- **RDF**: the original annotation format (RDF) produced by the NotaBene annotation tool. See the RDF/XML syntax specification and the NotaBene documentation for more information. The RDF file can be used to correct the annotation in NotaBene, but you need to pair it with the XML text source file. Please contact us if you want to work with the RDF files.

This list of texts is not final. Please refer to the project homepage http://srcmf.org for the up-to-date list.

Title	Date	Words
Serments de Strasbourg	842	117
Sequence de Sainte Eulalie	881	191
Vie de Saint Alexis	vers 1050	4 832
Passion de Clermont	2nde moitié 10e	2 749
Vie Saint Léger	2nde moitié 10e	1 398
Chanson de Roland	vers 1100	29 338
Lapidaire en prose	milieu 12e	4 799
Tristan de Beroul	entre 1165 et 1200	27 257
Yvain de Chretien de Troyes	1177-81	42 103
Quatre Livres des Rois	fin 12e	40 000
Aucassin et Nicolete	fin 12e/déb. 13e	10 009
La Conqueste de Constantinople de R. de Clari	après 1205	33 994
Queste del Saint Graal	vers 1220	40 000
Miracles de G. de Coinci	1218-1227	25 000
Roman de la Rose de J.de Meun	1269-1278	20 000

3 Documentation

3.1 First readings

- The SRCMF tutorial (PDF)

- The list of categories (English/French) and a tree view of the hierarchy of syntactic categories (1 page, PDF)

3.2 The SRCMF grammar model

The syntactic categories of the SRCMF annotation and the grammatical principles of the annotation are explained in detail in the Fiches de documentation des catégories SRCMF.

The *fiches* are a compilation of filecards (in French), with one card for each category. They contain a large number of annotation examples and are intended as a reference for corpus users and a guideline for the annotation of new texts.

3.3 Further reading

- Have a look at the presentation of the corpus at the LSRL 43 (New York, 2013) to get an idea of the corpus and the annotation principles. Download the slides (pdf) or watch the tools in action on YouTube (17min.).
- Use the TigerSearch manual (part of the TigerSearch installation, chapters 3 and 4) or the documentation on the TigerSearch homepage
- This TIGER/SRCMF tutorial for students includes hints for the TIGERSearch installation and more query examples.
- Or try some of the queries we prepared for the SRCMF annotation (see section 4.2.3 below).

4 Tools

4.1 NotaBene

NotaBene is a tool for syntactic corpus annotation, developed by Nicolas Mazziotta. It is described in Mazziotta (2010a).

- Download NotaBene on Sourceforge.
- The latest NotaBene documentation is included in the distribution.

4.2 TigerSearch

4.2.1 TigerSearch distribution

- TigerSearch Homepage. TigerSearch was developed by the Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, see Lezius (2002).
- Updated TigerSearch Distributions (ZIP files): use these files for the installation (the original windows installation file does not work in all versions of Windows, and the tiger.jar had to be updated for Max OS X).

- for Windows: ZIP file (42 MB), unpack and install under C: (i.e. in the top directory of your hard disk)
  Start the programme by clicking on the file
  `C:\TigerSearch\bin\TigerSearch.exe`
  (don't confound the exe file with the Tiger icon file)
- for Mac: ZIP file, 44 MB, unpack and install under [your home directory]/Applications
  Start the programme by clicking on the file
  `[your home directory]/Applications/TIGERSearch/lib/runTS.command`
  (if you wish to install the programme elsewhere, you must adapt the path in runTS.command)

4.2.2 Syntactic concordances for TigerSearch

The KNIC Concordances tool allows the user to transform TIGER XML files containing search results into a tabular format. See Rainsford & Heiden (2014) (pdf version) for a scientific presentation of the tool; and the tutorial contains instructions specific to SRCMF.

4.2.3 TigerSearch queries for SRCMF

Queries will be published here. In the meantime, consult the TIGER/SRCMF tutorial for some commented query examples (ready to paste into TigerSearch).

4.3 TXM

The text/corpus analysis platform TXM is available on Sourceforge for various platforms (Windows, Linux, Mac OS X).

4.4 TreeTagger and parameter files

The TreeTagger software can be downloaded from Helmut Schmid's website. The SRCMF project provides:

- a TreeTagger Parameter file for Old French with the Cattex 2009 tagset (without lemmatisation): from this BFM Website.
- a TreeTagger Parameter file for Old French with the NCA tagset, including partial lemmatisation: from Achim Stein's homepage (ZIP file). See the list of tags and Kunstmann & Stein (2007) for information on the NCA tagset.

4.5 Models for dependency parsing

Parser models made available here may be downloaded for non-commercial, non-profit use only.

- For **mate tools** parsers: *link to homepage* (Björkelund et al., 2010; Bohnet, 2010; Bohnet & Nivre, 2012).

– The LREC 2014 model (ZIP, 47MB) was trained on 11 SRCMF texts (206323 words). Input text must be tokenized as in SRCMF (spaces after apostrophes, e.g. *l' ot*, not *l'ot*). Punctuation is not recognized. The training procedure is described in Stein (2014).

– Updated Old French models for two mate tools parsers, trained on SRCMF 0.9 texts, as described in Stein (2016). The paper and a *readme* file are contained in the zip archives. These models are provided for the LREC *share your LRs* initiative.

  * for the mate tools joint transition parser (zip, 362MB): joint analysis of part of speech, morphological features, and dependencies
  * for the mate tools graph-based parser (zip 67MB): contrary to the approach described in the LREC paper, this package contains models for a full mate tools pipeline including lemmatisation, part of speech, morphological features, and dependencies. Results can be slightly improved if lemmatisation and tagging are done using *TreeTagger* and *Marmot*, see my other resources.

References

Björkelund, Anders, Bernd Bohnet, Love Hafdell & Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations* COLING '10, 33–36. Stroudsburg, PA, USA: Association for Computational Linguistics. http://portal.acm.org/citation.cfm?id=1944284.1944293.

Bohnet, Bernd. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 89–97. Beijing, China: Coling 2010 Organizing Committee. http://www.aclweb.org/anthology/C10-1011.

Bohnet, Bernd & Joakim Nivre. 2012. A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1455–1465. Jeju Island, Korea: Association for Computational Linguistics. http://www.aclweb.org/anthology/D12-1133.

Glikman, Julie. 2009. *Parataxe et Subordination en Ancien Français*. Paris, Potsdam: Université Paris Ouest Nanterre/Universität Potsdam.

Glikman, Julie. 2012. Les incises en croire, cuidier, penser en Ancien Français. *Linx. Numéro thématique 'Entre rection et incidence: des constructions verbales atypiques?'* https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxnbGlrbWFuanVsaWV8Z3g6NTM1OTU4ZWI4MzQ5YzM4OQ.

Glikman, Julie & Nicolas Mazziotta. 2011. Représentation de l'oral et structures syntaxiques dans la prose de la Queste del saint Graal. In *Représentation du sens linguistique V, 25-27 Mai 2011, Chambéry*, Chambéry: Éditions de l'Université de Savoie.

Guillot, Céline, Christiane Marchello-Nizia & Alexeij Lavrentiev. 2007. La Base de Français Médiéval (BFM): états et perspectives. In Pierre Kunstmann & Achim Stein (eds.), *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, Stuttgart: Steiner.

Heiden, Serge. 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*, .

Heiden, Serge & Alexei Lavrentiev. 2012. The TXM Portal Software giving access to Old French Manuscripts Online. In *Proceedings of the 1st Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects, Seventh International Conference on Language Resources and Evaluation (ELRA), Istanbul, Turkey*, .

Heiden, Serge, Jean-Philippe Magué & Bénédicte Pincemin. 2010. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In Sergio Bolasco, Isabella Chiari & Luca Giuliano (eds.), *Statistical Analysis of Textual Data-Proceedings of 10th International Conference JADT 2010, Rome, 9-11 juin 2010*, http://www.ledonline.it.

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

Lavrentiev, Alexei. 2010. La 'phrase' en français médiéval : une réalité ou une reconstruction artificielle? In Franck Neveu et al. (eds.), *Actes du 2e Congrès Mondial de Linguistique Française, La Nouvelle Orléans, 12-15 juillet 2010*, 277–289. Institut de Linguistique Française. http://dx.doi.org/10.1051/cmlf/2010125.

Lezius, Wolfgang. 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora (German)* University of Stuttgart Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), vol. 8, no. 4. Stuttgart: Institut für Maschinelle Sprachverarbeitung (IMS).

Mazziotta, Nicolas. 2009. *Ponctuation et syntaxe dans la langue française médiévale. Étude d'un corpus de chartes originales écrites à Liège entre 1236 et 1291* Beihefte zur Zeitschrift für romanische Philologie; 354. Tübingen: Niemeyer.

Mazziotta, Nicolas. 2010a. Building the 'Syntactic Reference Corpus of Medieval French' using NotaBene RDF Annotation Tool. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*, www.aclweb.org/anthology/W/W10/W10-1820.pdf.

Mazziotta, Nicolas. 2010b. Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples. *Recherches qualitatives. Hors-série 'Les actes'* 9. 83–94.

Mazziotta, Nicolas. 2012. Le Syntactic Reference Corpus of Medieval French: Structure, outils et exploitation. In A. Dister, D. Longrée & G. Purnelle (eds.), *11es Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 701–713. Liège, Bruxelles: Université de Liège.

Mazziotta, Nicolas. in press. Traitement de la coordination dans le Syntactic Reference Corpus of Medieval French (SRCMF). In *Actes du XXVIe Congrès de linguistique et de philologie romanes (València, 2010)*, Berlin: De Gruyter. acceptedforpublication15.12.2010.

Pincemin, Bénédicte, Serge Heiden, Marie-Hélène Lay, Jean-Marc Leblanc & Jean-Marie Viprey. 2010. Fonctionnalités textométriques : Proposition de typologie selon un point de vue utilisateur. In Sergio Bolasco, Isabella Chiari & Luca Giuliano (eds.), *Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010, Rome, 9-11 juin 2010*, http://www.ledonline.it.

Prévost, Sophie. 2002. Evolution de la syntaxe du pronom personnel sujet depuis le français médiéval: la disparition d'alternances signifiantes. In Dominique Lagorgette & Pierre Larrivée (eds.), *Représentations du sens linguistique*, 309–329. München: Lincom.

Prévost, Sophie. 2003. Détachement et topicalisation: des niveaux d'analyse différents. *Cahiers de praxématique* 40. 97–126.

Prévost, Sophie. 2009. Topicalisation, focalisation et constructions syntaxiques en français médiéval : des relations complexes. In D. Apothéloz, B. Combettes & F. Neveu (eds.), *Les linguistiques du détachement, actes du colloque international de Nancy*, 427–439. Bern: Peter Lang.

Prévost, Sophie. 2010. Évolution de la position du sujet pronominal en français médiéval: une approche sémantico-pragmatique. In Franck Neveu et al. (eds.), *Congrès Mondial de Linguistique Française - CMLF 2010*, Paris: Institut de Linguistique Française. http://dx.doi.org/10.1051/cmlf/2010106.

Prévost, Sophie. 2011a. *Expression et position du sujet pronominal du 12ème au 14ème siècle: une approche quantitative.* Paris: Recherche inédite en vue de l'obtention de l'HDR. http://tel.archives-ouvertes.fr/tel-00667183.

Prévost, Sophie. 2011b. Expression et position du sujet pronominal en français. In Jacques François & Sophie Prévost (eds.), *L'évolution grammaticale à travers les langues romanes. Mémoires de la Société de Linguistique de Paris, Tome XIX*, 13–33.

Rainsford, Thomas. 2011. *The Emergence of Group Stress in Medieval French.* PhD Thesis: University of Cambridge. http://www.dspace.cam.ac.uk/handle/1810/243900.

Rainsford, Thomas, Céline Guillot, Alexei Lavrentiev & Sophie Prévost. 2012. La zone préverbale en ancien français : apport des corpus annotés. In *Actes du 3e Congrès Actes du 3e Congrès Mondial de Linguistique Française (CMLF), Lyon, July 2012*, Paris: Institut de Linguistique française.

Rainsford, Thomas M. & Serge Heiden. 2014. Key Node in Context (KNIC) Concordances: Improving Usability of an Old French Treebank',. *SHS Web of Conferences* 8. 2707–2718.

Stein, Achim. 2008. Syntactic Annotation of Old French Text Corpora. *Corpus* 7. 157–171.

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). http://www.lrec-conf.org/proceedings/lrec2014/pdf/239_Paper.pdf.

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). http://www.uni-stuttgart.de/lingrom/stein/downloads/stein2016_

old-french-dependency-parsing.pdf.
Stein, Achim & Sophie Prévost. 2013. Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF). In Paul Bennett, Martin Durrell, Silke Scheible & Richard Whitt (eds.), *New Methods in Historical Corpora* Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3, 275–282. Tübingen: Narr.
Stein, Achim & Helmut Schmid. 1995. Étiquetage morphologique de textes français avec un arbre de décisions. *Traitement automatique des langues* Volume 36, Numéro 1-2: Traitements probabilistes et corpus. 23–35.
Stein, Achim & Carola Trips. 2012. Diachronic aspects of borrowing aspect: the role of Old French in the development of the 'be going to+INF' construction. In *Actes du 3e Congrès Mondial de Linguistique Française (CMLF), Lyon, 4-7 Juillet 2012*, 227–246. Paris: Institut de Linguistique française. http://dx.doi.org/10.1051/shsconf/20120100254.
Stein, Achim 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. http://www.uni-stuttgart.de/lingrom/stein/corpus/.

5 ChangeLog

21.10.13 Demo TIGER versions of two texts installed (graal, yvain). See Download section 2.2.

14.07.13 Internal TIGER version of the complete corpus installed. See Download section 2.2.

26.03.14 First parser models published (section 4.5). License conditions updated. (Internal) TIGER version of the complete corpus updated. New reviewed texts published.

02.03.15 Documentation files updated. (Internal) TIGER version of the complete corpus updated.

06.03.15 New texts published. License status changed for several texts.

24.09.15 License conditions changed for some texts. Version numbering for the full corpus introduced: this new version is 0.8. New attributes introduced to facilitate some Tiger queries (see documentation).

01.10.15 License conditions changed for Roman de la Rose.

25.10.15 New resources for dependency parsing.

13.11.15 Information about prose/verse added.

3.3.16 Word numbers in table updated. 'Quatre Livres des Rois' released. Direct speech marked in 'Yvain'. Full SRCMF corpus version 0.9. published.

11.3.16 SRCMF corpus version 0.91 published (bug fixes only in the full versions).

8.6.16 Old French parser models (presented at LREC 2016) published.

Email address: webmaster <at> <this domain>