Defining the maximal syntactic unit in the *Syntactic Reference Corpus of Medieval French:* Modeling, annotation, exploitation

Nicolas Mazziotta (Université de Liège)

This paper evaluates the consequences of the modeling choices made to annotate the corpus of the project *Syntactic Reference Corpus of Medieval French* (SRCMF; Stein & Prévost 2013; Mazziotta 2012). Theoretical choices are introduced (1) and two issues are surveyed (2-3). The objective of this paper is to provide an analytic retrospective evaluation of previous works by those who have used the resource and to provide a concrete suggestion (4).

1. Maximal syntactic structure. SRCMF aims at providing a comprehensive dependency-based analysis of microsyntactic relations (i.e. grammatical subordination and coordination relations) in a collection of texts selected from the *Base de Français Médiéval* (BFM, Guillot *et al.* 2007) and the *Nouveau Corpus d'Amsterdam* (Kunstmann & Stein 2007). Studies that focused on chunking medieval French texts into sentences (e.g. Mazziotta 2009 and Lavrentiev 2010) highlight the difficulties caused by the combination of morphosyntax, semantics and information structure (Hagège 1999) in this respect. Such a confusion leads to a poor consensus regarding the boundaries of the sentences – grammarians of the French language still emphasize this aspect; e.g. Wilmet 2003. In a project implying half a dozen of annotators with heterogeneous academic educations, it seems sensible to elaborate a list of non-ambiguous criteria to define the boundaries of a maximal syntactic unit (MaxS) rather than the ones matching the sentence in all its complexity. This MaxS is dubbed "maximal" for relations between MaxS's are posited not to be of microsyntactic nature. MaxS's therefore mainly correspond to independent clauses elaborated around a finite verb (http://srcmf.org/fiches/Max.html) and coordination between MaxS's is not a microsyntactic relation; consequently, arguments are never shared among several finite verbs. These modeling choices allow for a consistent annotation of the corpus. However, they convey inherent problems when it comes to using it.

2. Subject and shared arguments. In medieval French clauses do not have to include an external subject. SRCMF made possible to investigate the preverbal complements in medieval French in a perspective that implies count data and statistics (Rainsford *et al.* 2012). The methodology of data extraction relies on an external tool that matches most of the features of the corpus: TigerSearch (http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/tigersearch.en.html). As many other syntactic exploration tools, this tool cannot access to relations between MaxS's either. In cases such as: *Et il gite la main et prend l'espee par le heut...* "An he throws his hand and takes the sword by the hilt" (BFM: qgraal 161b, 3) (where *gite* and *prend* share a unique subject), automatic processing simply outputs that *prend* has no external subject. Since MaxS's cannot share any argument, some of the results of automated extractions must be manually processed.

3. Represented orality (RO). Studies about the syntactic specificities of so-called *represented orality* (i.e. written text given as a representation of oral language, such as quotations) suffer from the same kind of restrictions (s. a.o. Glikman et Mazziotta 2013), but this field of investigation also raises its own difficulties. Rules for identifying MaxS's applied to RO lead to two distincts cases: (i) either the *verbum dicendi* (VD) is placed at the beginning or at the end of the quotation; (ii) or it is used as an insert. In the first case, the annotation encodes two separate MaxS's (cf. Tesnière 1959 ; Wilmet 2003), even if the order is VS ; in the second case, the insert is identifed as a dependent of the quoted clause. However, VD have similar morphosyntactic characteristics to other verbs in RO (Marchello-Nizia 2012). Without a reliable lemmatization, it remains impossible to automatize the distinction between VD at the boundaries of RO (case (i) above) and other MaxS's entirely pertaining to the narrative apparatus.

4. Suggestion. SRCMF uses graph-like RDF conventions to encode the analysis, which makes it possible to add distinct layers of annotation in a modular way. As far as the aforementioned problems are concerned, an additional layer describing argument sharing (fig. 1) and relations between VD and OR (fig. 2) can be added easily without changing the structure of the original SRCMF annotation. Supplemental layers can be left aside to preserve current exploration

procedures that use a tree-like model (TigerSearch's or CoNLL's, s. http://www.conll.org/) or they can be directly exploited using graph-aware tools (such as SPARQL engines).

Figures

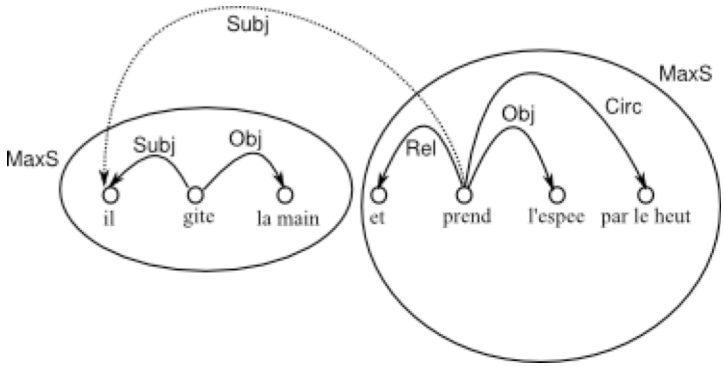

Figure 1: Argument encoding – The dotted edge is of the same (microsyntactic) kind as other edges in the figure, but it is part of another graph. This graph that can be merged (thus breaking planarity as well as the tree-object constraint stating that each node must have only one parent) with the original ones including each MaxS or remain separated.

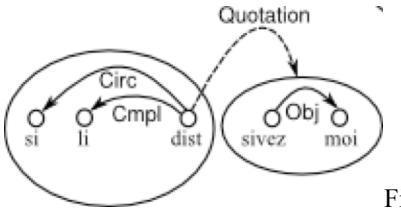

Figure 2: Quotation relations – The dashed edge is not of the same kind as other edges : it represents a discursive link between the two MaxS's and links the VD to the utterance it introduces (graphs actually need to undergo a formal transformation to allow for structures to be represented as vertices of an edge ; s. Kahane/Mazziotta 2015).

References

Glikman, Julie et Mazziotta, Nicolas (2012). *Représentation de l'oral et syntaxe dans la prose de la Queste del saint Graal (1225-1230)*. In Lagorgette, D. & Larrivée, P. (edd.). Représentations du sens linguistique V, Chambery : Éditions de l'Université de Savoie.

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

Hagège, Claude. 1999. *La structure des langues*. Paris: PUF.

Kahane, Sylvain & Mazziotta, Nicolas. 2015. Syntactic polygraphs. A formalism extending both constituency and dependency. In *Proceedings of MOL 2015*.

Kunstmann, Pierre & Stein, Achim. 2007. Le Nouveau Corpus d'Amsterdam. In Pierre Kunstmann & Achim Stein (edd.), *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.

Marchello-Nizia, Christiane. 2012. L'oral représenté: un accès construit à une face cachée des langues 'mortes'. In Guillot C., Combettes B., Lavrentiev A., Oppermann-Marsaux E. & Prévost S. (edd.). *Le changement en français. Études de linguistiquediachronique*. Bern/Berlin/Bruxelles: Peter Lang.

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*. Tübingen: Niemeyer.

Prévost, Sophie. 2016. *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. .
Rainsford, Thomas, Guillot, Céline, Lavrentiev, Alexei et Prévost, Sophie. 2012. La zone préverbale en ancien français : apport de corpus annotés. *In Actes du 3e Congrès Mondial de Linguistique Française,* 159-176.
Stein, Achim & Prévost, Sophie. 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.
Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.
Wilmet, Marc. 2003. Grammaire critique du français. Bruxelles: Duculot.