Diachronic syntax based on constituency and dependency annotated corpora: theoretical and methodological issues.

Achim Stein (University of Stuttgart)
achim.stein@ling.uni-stuttgart.de

This paper introduces a new, dependency-annotated corpus of Medieval French, the *Syntactic Reference Corpus of Medieval French (SRCMF)*[1] and compares it to the Penn-style phrase structure annotated corpus *Modéliser le changement: les voies du français (MCVF)*[2].

In the first part, I will introduce the SRCMF annotation model. In the second part, I will present a case study of constructions with information-structural (IS) properties like clefts and dislocations and will serve to discuss two more general and related problems:

1. How can we use corpus data to induce properties of historical stages of languages when ambiguities on several linguistic levels cannot be directly resolved?
2. What is the contribution of the corpus annotation model to this task? How can it reconcile the wish to be maximally expressive while being minimally interpretive?

Take a look at the following French cleft constructions (Prévost, 2009, 3):

(1) C'est Paul qui est tombé? Non, **c'est Luc qui est tombé**.
is it Paul who has fallen? no it is Luc who has fallen.

(2) Qu'est-ce qui se passe? **C'est Luc qui est tombé**.
what is it that REFL happens? it is Luc who has fallen.

Some authors capture the difference between (1) and (2) by distinguishing several types of clefts: cases like (1) have been called stressed-focus clefts, cases like (2) informative-presupposition clefts (for English classifications see e.g. Prince 1978; Collins 1991, for French Dufter 2008). The ambiguity resides in the IS status of *Luc* and pertains to different linguistic levels:

(3)
- Reference: the pronoun *ce* has the status of either an expletive or a demonstrative
- Semantics: the verb *est* is either a full copula verb or a grammaticalized IS-marker
- Information structure: the XP is either a predicative complement or IS-marked. If it has the latter status it is either a focus or a topic
- Syntax: the subordinate clause depends either on the XP or on the pronoun
- Prosody: the XP is either prosodically marked or not

Knowledge about the properties of Modern French can help to disambiguate these constructions: (1) it is a strict SVO language, (2) it has no sentence-initial focus position, (3) stress is syntactically constrained to the end of prosodic groups, (4) it shows neither null subjects nor Verb Second. For diachronic stages of French, this knowledge is not available because these properties have undergone a number of drastic changes, so that variation of all four of them was characteristic for the Old French period (OF: 842-ca.1320):

(4) 1. word order was changing from SOV to SVO;

[1] http://srcmf.org (Prévost and Stein, 2012).

[2] http://www.voies.uottawa.ca/corpus_pg_fr.html (Martineau, 2009)

2. OF still had a sentence-initial focus position;
3. OF still showed rather unconstrained word accent;
4. OF still had null subjects as well as Verb Second orders.

We also cast a glance on the better established Old English (OE) and Middle English Penn corpora (Taylor et al., 2003; Kroch and Taylor, 2000), where similar problems arise since from late OE on the change from SOV to SVO, the loss of null subjects and Verb Second resulted in new information-structural possibilities. The problem concerning clefts is almost analogous to Medieval French, since *(h)it* can be a personal or an expletive pronoun, the demonstrative *þat* can also introduce a cleft, *beon/ben* is either a full copula verb or a grammaticalized IS-marker etc.

It is obvious that in such a situation annotators of historical corpora face serious problems: in the case of clefts, syntactic annotation can either limit itself to syntactic structure (here: a predicative main clause with pronominal or null subject and a relative subordinate clause) or provide a functional interpretation by marking the same structure with a "cleft tag" (CP-CLF in the Penn annotation). The latter option, however, presents a historical datum (a structure) as if it were evidence for a given phenomenon (a cleft), and doing this is tantamount to stating that the multiple ambiguities in (3) were properly evaluated, not only in the immediate context, but also in the context of the language specific variations (4).

For reasons of time, previous corpus-based studies (e.g. Bouchard et al. 2010 and earlier papers) can only be considered marginally; the main interest is technical: we will show how the corpora MCVF and SRCMF handle the aforementioned problems and discuss them from the users' and the annotators' perspective.

In the final part, I will show that although the choice of the syntactic categories was motivated by linguistic considerations rather than by parsing efficiency, the SRCMF dependency model can be used successfully with state-of-the-art parsers like *mate* (Björkelund et al., 2010) for the automatic annotation of Old French texts.

Björkelund, A., Bohnet, B., Hafdell, 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, 33–36. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bouchard, J., Dufresne, M. and Dupuis, F. 2010. "Les changements dans les constructions à copule et évolution des clivées en français et en anglais médiéval". In *Le changement en français: études de linguistique diachronique*, B. Combettes (ed), volume 89 of *Sciences pour la communication*, 73–91. Bern: Lang.

Collins, P. 1991. *Cleft and pseudo-cleft constructions in English*. London, New York: Routledge.

Dufter, A. 2008. "On explaining the rise of c'est-clefts in French". In *The Paradox of Grammatical Change*, U. Detges and R. Waltereit (eds), 31–56. Amsterdam, Philadelphia: Benjamins.

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

Martineau, F., (ed). 2009. *Le corpus MCVF. Modéliser le changement: les voies du français*. Ottawa: Université d'Ottawa.

Prévost, S. 2009. "Topicalisation, focalisation et constructions syntaxiques en français médiéval : des relations complexes". In *Les linguistiques du détachement, actes du colloque international de Nancy*, B. C. e. F. N. D. Apothéloz (ed), 427–439. Bern: Peter Lang.

Prévost, S. and Stein, A., (eds). 2012. *Syntactic Reference Corpus of Medieval French*. Lyon/Stuttgart: ENS de Lyon/Universität Stuttgart.

Prince, E. 1978. "A comparison of wh-clefts and it-clefts in discourse". *Language* 54: 883–906.

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