Coreference annotation with SACR, a new drag-and-drop based tool

Bruno Oberle

Coreference annotation with SACR, a new drag-and-drop based tool

Bruno Oberle

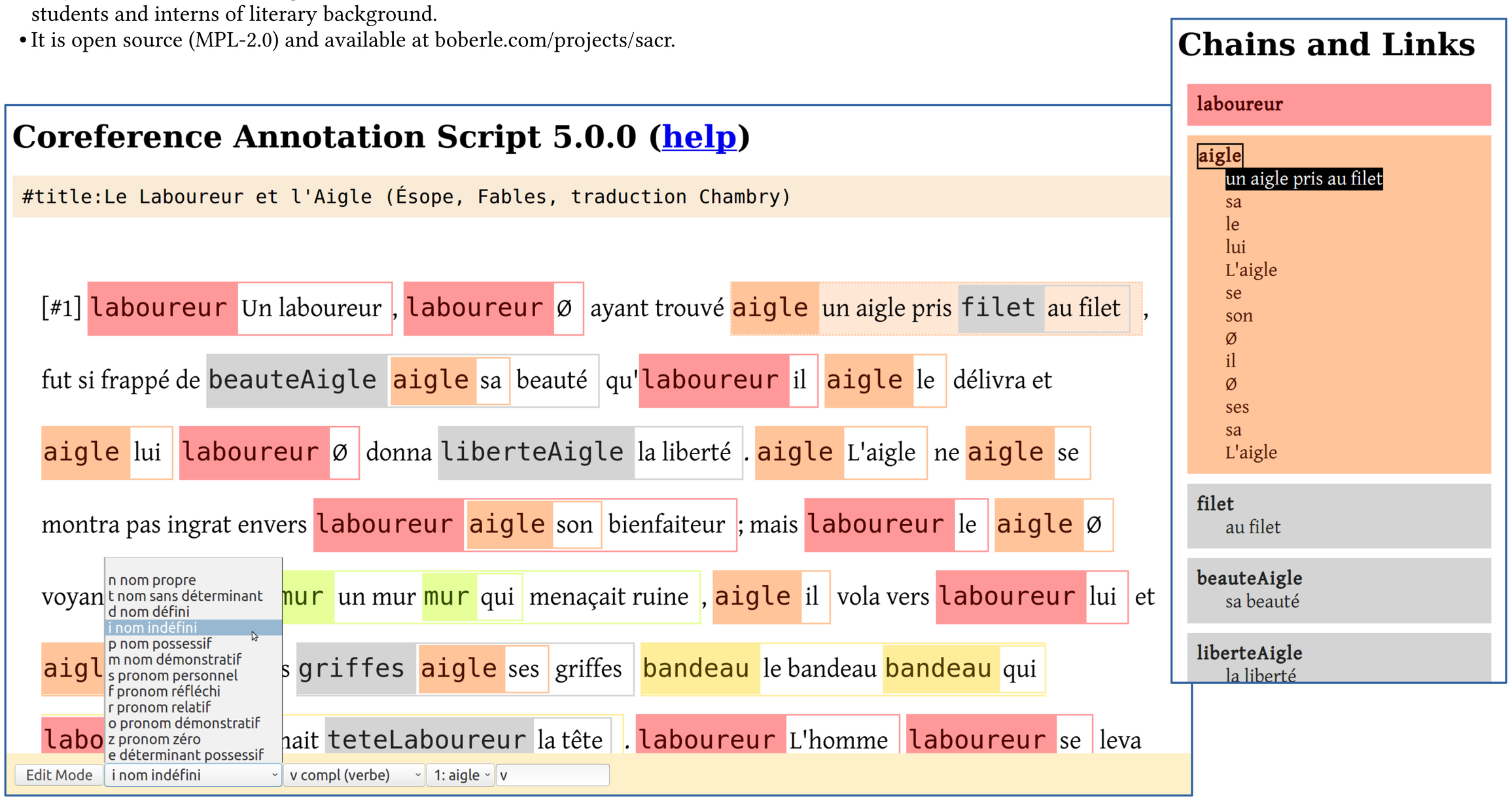
LiLPa, Université de Strasbourg

SACR and its context

- NLP applications need large, manually annotated corpora but careful annotation is time consuming, so annotation tools must require the least effort from the annotator.
- Furthermore, annotation is often done by students, interns or non-technical users, so annotation tools must be ready-to-use and have an intuitive interface.
- SACR ("Script d'Annotation des Chaînes de Référence") is implemented as a simple web page (HTML and Javascript) running in Firefox and Chrome: it is thus cross-platform.
- Easy to use, with no learning time and no installation, it is well-suited for students and interns of literary background.
- It is open source (MPL-2.0) and available at boberle.com/projects/sacr.

Coreference annotation with SACR

1. Delimiting and marking referring expressions (*i.e.* linguistic expressions that refer to some extra-linguistic entity) is done by clicking on the first and last tokens of the expression.
2. Linking two coreferential expressions is done by drag-and-dropping the expressions one over the other.
3. Feature annotation (part of speech, grammatical function, etc.) is done by selecting an attribute from a list or by using shortcut keys (like "d" for a noun with a definite article or "r" for a relative pronoun): the program then goes automatically to the next expression to be annotated



Features to help the annotator

- A list of chains and their elements where the annotator can access all the annotations and drag-and-drop them from or to the main window.
- A tool to search elements by feature (*e.g.* all nouns and/or pronouns, all subjects, etc.) or to check that no annotation has been forgotten.

Export and exchange formats

- Annotations are stored in an easy-to-parse text format.
- Scripts allow to convert from or to Glozz (Widlöcher & Mathet 2012), TXM (Heiden 2010), Analec (Landragin, Poibeau & Victorri 2012) and the CONLL-2011 format.
- SACR can be used alongside automatic annotation tools like a chunker or a part of speech tagger, allowing a succession of manual and automatic steps, as used in the Democrat project (Poudat & Landragin 2017).

References

- Heiden S. (2010). The TXM Platform: Building Open Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Ryo Otoguro, *et al.* (eds.), *24th Pacific Asia Conference on Language, Information and Computation.* Sendai, Japan. Wasea University.
- Landragin F., Poibeau T. and Victorri B. (2012). Analec: a new tool for the dynamic annotation of textual data. In Nicoletta Calzolari *et al.* (eds). *Proceedings of LREC'12.* Istanbul, Turkey. ELRA.
- Poudat C. and Landragin F. (2017). *Explorer un corpus textuel.* Champs linguistiques. De Boeck.
- Widlöcher A. and Mathet Y. (2012). The Glozz Platform: A corpus annotation and mining tool. In *Proceedings of the 2012 ACM symposium on Document engineering.*

Acknowledgement

This work was supported by the Democrat project "DEscription et Modélisation des Chaînes de Référence: outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique" *(Description and modeling of reference chains: tools for corpus annotation (with diachronic and cross-linguistic approaches) and automatic processing)*, ANR-15-CE38-0008.