The Journal of Open Source Software

Samewords: Automatic word disambigation in critical text editions

Michael Stenskjær Christensen[1,2]

1 Saxo-Institute, University of Copenhagen **2** Representation and Reality, University of Gothenburg

DOI: 10.21105/joss.00941

Software

- Review
- Repository
- Archive

Submitted: 05 July 2018
Published: 09 April 2019

Summary

The presented software helps the editor of critical scholarly editions solve a common problem when presenting a critical apparatus. To illustrate the purpose and use of the software, I will describe the scholarly domain it applies to.

Scholarly disciplines that use texts as their primary source of data are confronted with the challenge of textual editing. This applies to virtually all studies within the arts and humanities and certainly some that lie beyond; a few examples would be literary studies, philosophy, history, and linguistics. Textual, or scholarly, editing is the activity of providing texts that will be the basis for any subsequent work, be it analysis by a researcher, translation, or leisurely reading. This activity thus provides the raw material for any kind of research within these disciplines.

Establishing such texts is a highly specialized activity presupposing research skills in all the relevant subject areas of a text, such as the linguistic disciplines, the study of the content, but also the whole discipline of stemmatics, codicology, palaeography (it very often involves reading medieval or early modern material in manuscript form), and text history. The result of this work is the critical text, a text which presents the closest an editor is able to get to the ideal text defined by the use case and purpose of the edition.

The critical apparatus is an essential part of a critical text. This gives any expert reader information about the choices that the editor has made in establishing the text. Examples of such information could be the differences between the text of different witnesses (for example manuscripts or printings), details of the history of the text, and particular difficulties in the interpretation of the witnesses.

A common problem for the editor, and later a reader, of a critical text edition is the handling of ambiguous references in the critical apparatus. Let us take this paragraph as an example:

```
Here is a chunk of text, what a nice place for a critical note.

----
1 a] om. M
```

Unless the “a” is disambiguated, it is impossible to determine which instance the reference points to. This will often be done by a numbering scheme such as this:

```
Here is a chunk of text, what a nice place for a critical note.

----
1 a²] om. M
```

Reledmac (Rouquette & Wilson, 2018) is the standard LaTeX package used for typesetting critical scholarly editions of the highest standard. It already provides facilities for disambiguating identical words, but it requires the editor of the critical text to mark all potential instances of ambiguous references manually. This is a significant labour in large text editions, as any recompilation may change the presentation of the text, and hence require the editor to check for any new conflicts and annotate them accordingly. The annotation of ambiguous words can also be very complex, and the manual annotation therefore includes a large risk of error.

Samewords therefore automates this process. It is a Python 3 package that can be installed via `pip`, but an online interface and API is also provided for the users who are not used to installing and running software from the command line. It provides full Unicode 10 support, and handles single word conflicts by default (with the option to annotate multi-word conflicts) as well as apparatus entries with custom lemmas. It is possible to indicate custom ellipsis patterns for spans in custom lemma references. Further details, such as the number of context words to compare, recognized punctuation characters, and case sensitivity can be configured in a configuration file.

The source code has been archived at *Zenodo* with the linked DOI: (Christensen, 2018). The full documentation can be found at https://samewords.readthedocs.io/en/latest/.

Acknowledgements

I acknowledge valuable contributions from Florian Grammel who has reported numerous bugs and performed extensive testing and feedback.

References

Christensen, M. S. (2018, July). stenskjaer/samewords: Word disambiguation in critical text editions. doi:10.5281/zenodo.1306293

Rouquette, M., & Wilson, P. R. (2018). Reledmac: Typeset scholarly editions with latex. Retrieved from https://www.ctan.org/pkg/reledmac