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.00810

Software

- Review
- Repository
- Archive

Submitted: 05 July 2018
Published: 05 August 2018

Summary

A common problem for the editor of a scholarly textual 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 and 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The source code has been archived at *Zenodo* with the linked DOI: (Christensen 2018).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, Michael Stenskjær. 2018. "stenskjaer/samewords: Word disambiguation in critical text editions." https://doi.org/10.5281/zenodo.1306293.

Rouquette, Maïeul, and Peter R. Wilson. 2018. "Reledmac: Typeset Scholarly Editions with Latex," May. CTAN. https://www.ctan.org/pkg/reledmac.