



Multilingual Workflows for Semantic Change Research

PAOLA MARONGIU 

BARBARA MCGILLIVRAY 

ANAS FAHAD KHAN 

*Author affiliations can be found in the back matter of this article

COLLECTION:
DATA AND
WORKFLOWS FOR
MULTILINGUAL
DIGITAL HUMANITIES

DATA PAPER

]u[ubiquity press

ABSTRACT

We present a series of workflows that aim to support research in lexical semantic change, i.e. the phenomenon by which words change their meaning over time. The workflows each consist of a series of steps required to detect words that have undergone semantic change as evidenced by a corpus and cover a range of user scenarios, including lexicology, historical research, and legal studies. The workflows were created following the model adopted by the Social Sciences and Humanities Open MarketPlace and were designed in the context of a project on CLARIN resource families supported by CLARIN ERIC. In this paper, we present a use case for the workflow on lexicology, referring to resources for Latin and ancient Greek.

CORRESPONDING AUTHOR:

Paola Marongiu

Institut des sciences du langage (ISLa), University of Neuchâtel, Neuchâtel, Switzerland

paola.marongiu@unine.ch

KEYWORDS:

semantic change; historical languages; workflows; humanities and social sciences; open marketplace

TO CITE THIS ARTICLE:

Marongiu, P., McGillivray, B., & Khan, A. F. (2024). Multilingual Workflows for Semantic Change Research. *Journal of Open Humanities Data*, 10: 15, pp. 1–5. DOI: <https://doi.org/10.5334/johd.179>

(1) OVERVIEW

REPOSITORY LOCATION

The description of the workflows is part of an extended report about the outcomes of the CLARIN-funded project “A new CLARIN Resource Family for lexical semantic change research”. The report and the workflows are stored in a Zenodo repository with DOI [10.5281/zenodo.8156199](https://doi.org/10.5281/zenodo.8156199).

The workflow for lexical semantic change research in lexicology is stored in the Social Sciences and Humanities (SSH) Open Marketplace as a “Workflow” item. The workflow is called “Semantic change analysis for lexicological studies”, and it is available at <https://marketplace.sshopencloud.eu/workflow/yStoh2>.

CONTEXT

The evolution of word meanings (lexical semantic change) is a highly relevant subject for linguists, lexicographers, and scholars across the humanities. Take the word “snowflake”: this word initially referred to ice crystals but has evolved to describe unique individuals or those (perceived to be) easily offended, reflecting shifts in culture and society.

In recent years, Natural Language Processing researchers have developed algorithms that predict lexical semantic change (Schlechtweg et al. 2020). Moreover, annotated texts provide valuable context for word meanings, and some researchers have proposed models that represent semantic change information contained in lexical resources like dictionaries as linked data (Armaselu et al. 2022). However, accessing resources and tools for semantic change research remains challenging due to fragmentation across various domains.

To address this challenge, the CLARIN-funded project “A new CLARIN Resource Family for lexical semantic change research” (McGillivray et al. 2023) aimed to centralise essential resources for semantic change research, encompassing datasets with word sense annotations, word embeddings derived from diachronic corpora, automated algorithms for detecting semantic change, and lexical resources.

As part of this project, we have developed workflows inspired by the SSH Open Marketplace.¹ These workflows guide users through a series of manageable steps, connecting each step with relevant datasets, manuals, and digital tools. We chose the SSH Open Marketplace as a model given its broad scope across the whole of the SSH and its robust infrastructure, which is supported by three prominent SSH European Research Infrastructures consortia, CESSDA,² CLARIN,³ and DARIAH.⁴ The workflows that we have developed in the context of this project are designed not only to apply to various research domains but also to be language-independent. The SSH Open Marketplace discovery platform enables linking to various resources for each step and even at the workflow level itself, provided that the resource in question is part of the platform. In this way, any resource for any language, as long as it is available on the platform, can be linked to the different steps.

(2) METHOD

The SSH Open Marketplace website offers guidance on the creation of workflows⁵, with the following steps: identification and description of workflows and the association of metadata with them; description of the various steps involved in the workflow; and finally, the identification of relevant resources. In our case, the previous research carried out by the first two authors in semantic change detection, together with a knowledge of the current literature on the topic,

¹ SSH Open Marketplace. Available at <https://marketplace.sshopencloud.eu/> (Last accessed 13 December 2023).

² Consortium of European Social Science Data Archives. Available at <https://www.cessda.eu/> (Last accessed 13 December 2023).

³ Common Language Resources and Technology Infrastructure. Available at <https://www.clarin.eu/> (Last accessed 13 December 2023).

⁴ Digital Research Infrastructure for the Arts and Humanities. Available at <https://www.dariah.eu/> (Last accessed 13 December 2023).

⁵ SSH Open Marketplace. How to create a workflow in the SSH Open Marketplace? Available at <https://marketplace.sshopencloud.eu/workflow/hmGpmv> (Last accessed 13 December 2023).

directly inspired the selection of the specific workflows we chose to develop. Compared to existing workflows particularly focussed on the NLP processing steps (LSCDetection;⁶ Gruppi et al. 2022), we were interested in covering, as far as possible, the information needs of the various disciplines which might potentially make use of the results of semantic change detection. Once we had a first draft of our workflows, we contacted numerous experts, choosing among the authors of the articles we consulted during the initial literature review process (references to the articles are provided with the workflows in Zenodo). We asked them to give feedback on the workflows based on their own experiences in the field. In our case, we also endeavoured to include references to CLARIN Resource Families (CRFs) in our workflows and base our workflows mainly around the different categories of resources featured in the CRFs. The intention here was to make our workflows as interoperable as possible by citing categories of resources hosted by one of the major European SSH infrastructures, which could potentially be a point of reference throughout the different SSH disciplines. These categories, moreover, are periodically updated with new resources.

Figures 1 and 2 show the workflow “Semantic change analysis for lexicological studies” that we created within the SSH Open Marketplace. Each workflow is introduced by describing the task and the type of audience targeted. A list of steps follows (Figure 1), which the user can expand to find additional information and examples of resources available on the platform that can be used to complete each specific step (Figure 2).



Workflow steps (6)		
1	Set up a corpus	Expand ▾
2	Split the corpus into different spans	Expand ▾
3	Train word embeddings on different spans	Expand ▾
4	Evaluation against a Gold Standard	Expand ▾
5	Qualitative analysis of the results of the word embeddings	Expand ▾
6	Additional analysis: qualify the type of semantic change	Expand ▾

Figure 1 Overview of the steps of the workflow “Semantic change analysis for lexicological studies” in the SSH Open Marketplace.

1 Set up a corpus Collapse ▲

The first step for the lexicologist is to set up a corpus (a collection of texts) to perform their analysis on a reference corpus for the target language. Such a corpus should capture variation on the axes that the researcher intends to study e.g. it should be diachronic, or contain texts belonging to different genres or disciplines and this information should ideally be available via the metadata describing each of the texts in the corpus. This corpus does not necessarily need to be sense-annotated, but it should be lemmatised and Part-of-Speech tagged to reduce data sparsity.

Related items (2)

 LatinISE corpus (version 4) This corpus consists of Latin texts from the 2nd century B.C. to the 21st century. Non-linguistic metadata include information on genre, title, century and specific date. The corpus is availa... Read more	 The Diorisis Ancient Greek Corpus This corpus consists of 820 texts spanning between the beginnings of the Ancient Greek literary tradition (Homer) to the fifth century AD. The texts are sourced from the Perseus... Read more
---	---

[Show more](#)

Figure 2 Step 1 of the workflow “Semantic change analysis for lexicological studies”, as it appears in the SSH Open Marketplace.

⁶ **Garrafao** LSC Detection. Available at <https://github.com/Garrafao/LSCDetection> (Last accessed 12 December 2023).

(3) DATASET DESCRIPTION

OBJECT NAME

“Workflows for lexical semantic change” (in Zenodo); “Semantic change analysis for lexicological studies” (in SSH Open Marketplace).

FORMAT NAMES AND VERSIONS

The workflows stored in Zenodo are presented as a pdf file that follows the structure suggested by the SSH Open Marketplace. The workflow in the Open Marketplace is available as an online resource on the platform.

CREATION DATES

Start date: 2022-11-01

End date: 2023-10-31

DATASET CREATORS

Barbara McGillivray (Department of Digital Humanities, King’s College London, London, United Kingdom); Paola Marongiu (Institut des sciences du langage (ISLa), University of Neuchâtel, Neuchâtel, Switzerland); Fahad Khan (Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Pisa, Italy).

LANGUAGE

English, Latin, ancient Greek

LICENSE

Creative Commons Attribution 4.0 International

REPOSITORY NAME

Zenodo; Social Sciences and Humanities Open Marketplace <https://marketplace.sshopencloud.eu/>

PUBLICATION DATE

2023-10-31

(4) REUSE POTENTIAL

We have presented workflows that streamline and facilitate research in semantic change by simplifying access to relevant language resources and tools scattered across different repositories and platforms. The workflows are intended for use by a range of researchers from different backgrounds whose interests lie in the exploration of lexical semantic change as a valuable tool in addressing their respective research questions, whether in history, linguistics (including lexicology, lexicography and semantics), or cultural studies (McGillivray et al. 2023). The particular example of the lexicology workflow can be used for any study that aims to analyse the evolution of a lexical or semantic field, either to test specific hypotheses or to relate language changes to large-scale social and historical events. Beyond research, the workflows can be employed in teaching and learning scenarios, offering students an opportunity to engage with advanced linguistics concepts and data analysis techniques and exposing them to interdisciplinary approaches to the study of semantic change.

However, some challenges may arise from the reuse of the workflows in research contexts. Research in NLP on semantic change is evolving rapidly, requiring frequent updates to the more computational components of the workflows. From a more general perspective, workflows have been designed to be adaptable to the disciplines mentioned earlier. While this ensures comprehensive coverage of various research scenarios, this lack of granularity inevitably leads to overlooking some intersteps that are more specific to those disciplines and/or research questions. For instance, individual languages may lack some of the tools and resources

assumed by a single workflow, and it may be necessary to either build them or use resources that are available. We have chosen to forego granularity to ensure broader applicability of the workflows, fully aware that in the current state this may also represent a limitation to our work.

ACKNOWLEDGEMENTS

We wish to thank the experts Marton Ribary, Sandeep Soni, Lauren Klein, Jacob Eisenstein, Dani Roytburg and Emily Bell who provided their feedback on the workflows concerning their fields of expertise. We would also like to thank CLARIN-ERIC (especially Francesca Frontini) and SSHOC (especially Laure Barbot) for their support in realising this project.

FUNDING INFORMATION

This work was funded by CLARIN-ERIC through the CLARIN Resource Families project ‘A new CLARIN Resource Family for lexical semantic change research’ from November 2022 to June 2023.

COMPETING INTERESTS

Barbara McGillivray is editor-in-chief of the *Journal of Open Humanities Data* but did not take part in the editorial process or decisions pertaining to this manuscript.

AUTHOR CONTRIBUTIONS

PM: Conceptualization, Writing – original draft, Writing – review & editing

BMcG: Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing

AFK: Conceptualization, Supervision, Writing – original draft, Writing – review & editing

AUTHOR AFFILIATIONS

Paola Marongiu  orcid.org/0000-0002-5060-3307

Institut des sciences du langage (ISLa), University of Neuchâtel, Neuchâtel, Switzerland

Barbara McGillivray  orcid.org/0000-0003-3426-8200

Department of Digital Humanities, King's College London, London, United Kingdom

Anas Fahad Khan  orcid.org/0000-0002-1551-7438

Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche, Pisa, Italy

REFERENCES

- Armasele, F., Apostol, E.-S., Khan, A. F., Liebeskind, C., McGillivray, B., Truică, C.-O., Utka, A., Valūnaitė Oleškevičienė, G., & van Erp, M. (2022). LL(O)D and NLP Perspectives on Semantic Change for Humanities Research. *Semantic Web*, 13(6), 1051–1080. DOI: <https://doi.org/10.3233/SW-222848>
- Gruppi, M., Adali, S., & Chen, P.-Y. (2022). The SenSE Toolkit: A System for Visualization and Explanation of Semantic Shift. *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, in *Proceedings of Machine Learning Research*, 176, 283–287. Available at <https://proceedings.mlr.press/v176/gruppi22a.html>
- McGillivray, B., Khan, F., & Marongiu, P. (2023). A new CLARIN Resource Family for lexical semantic change – Final report. *Zenodo*. DOI: <https://doi.org/10.5281/zenodo.8156200>
- Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). Semeval-2020 task 1: Unsupervised lexical semantic change detection. In: A. Herbelot, X. Zhu, A. Palmer, J. Schneider, J. May & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1–23). Barcelona: International Committee for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2020.semeval-1.1>

TO CITE THIS ARTICLE:

Marongiu, P., McGillivray, B., & Khan, A. F. (2024). Multilingual Workflows for Semantic Change Research. *Journal of Open Humanities Data*, 10: 15, pp. 1–5. DOI: <https://doi.org/10.5334/johd.179>

Submitted: 01 November 2023

Accepted: 17 December 2023

Published: 30 January 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.