DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages

Dominik Schlechtweg,[♣] **Nina Tahmasebi,**[♠] **Simon Hengchen,**[♠] **Haim Dubossarsky,**[♦] **Barbara McGillivray**[◇,♡]

[♣]University of Stuttgart, [♠]University of Gothenburg, [♦]University of Cambridge, [◇]King's College London [♡]The Alan Turing Institute

Introduction

- traditional approach to annotate word senses are binary assignments to sense descriptions [Kilgarriff 1998]
  - ignores gradedness of word meaning
- two alternatives proposed by Erk et al. [2013]: [Erk et al. 2013]
  (i) graded judgments of word usage pairs (usage-usage)
  (ii) graded assignments of word usages to sense descriptions (usage-sense)
- judgments populate weighted graph [McCarthy et al. 2016]
- senses are not annotated directly, but **inferred** on the graph
- problems: applicability, scalability

Clustering

- correlation clustering [Bansal et al. 2004]
- optimization criterion: **minimize (weighted) number of cluster-edge conflicts** [Schlechtweg et al. 2020]

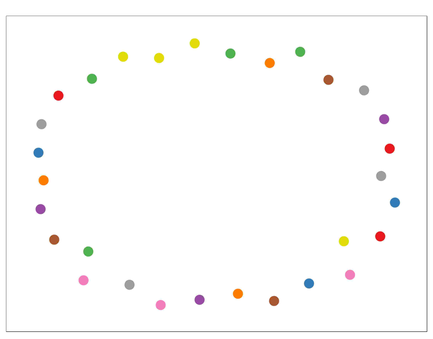
$$\arg\min_{C} L(C) = \sum_{e \in \phi_{E,C}} W'(e) + \sum_{e \in \psi_{E,C}} |W'(e)|$$

(i) finds the optimal number of clusters on its own
(ii) handles missing information (non-observed edges)
(iii) robust to errors by using the global information
(iv) respects the gradedness of word meaning
(v) dominated in simulation study

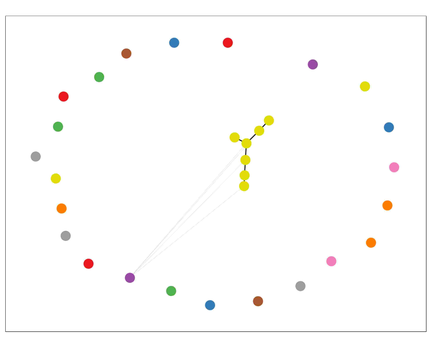
↑ 4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated

DURel relatedness scale.

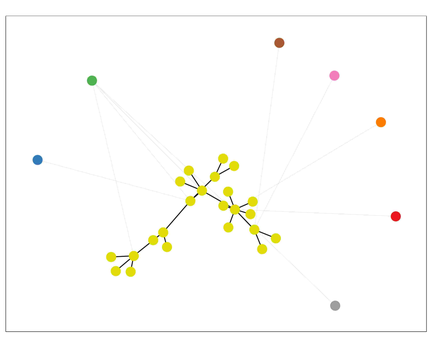
Edge Sampling



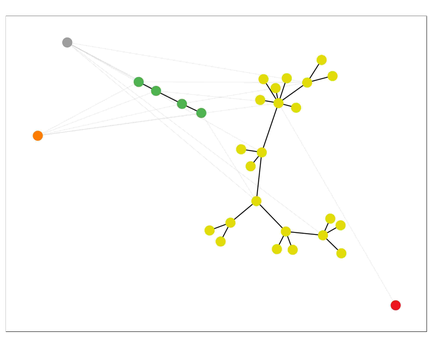
No information



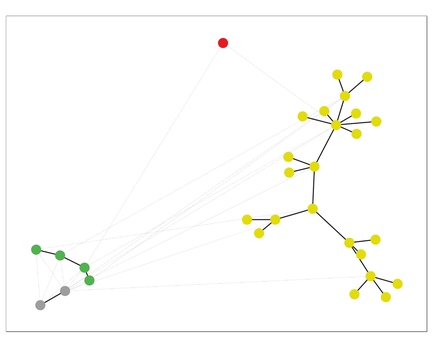
Initial clustering (exploration)



Cluster comparison (combination)



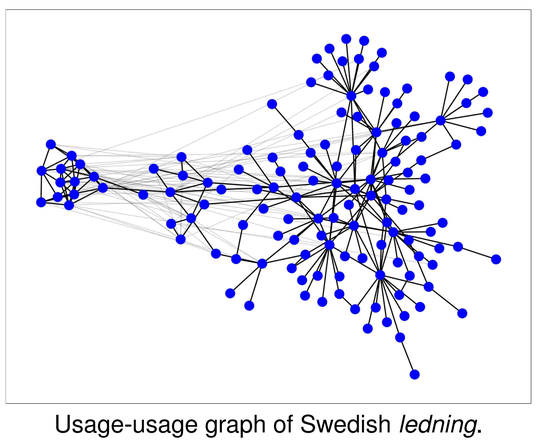
Compare non-assignable uses (exploration)



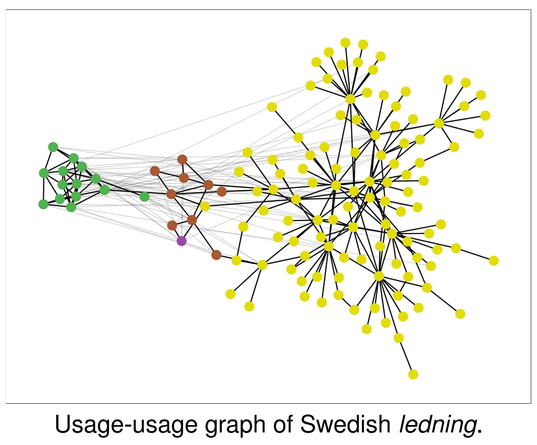
Combination and exploration

Procedure (i): Usage-Usage Graphs

(Usage) Von Hassel replied that he had such faith in the **plane** that he had no hesitation about allowing his only son to become a Starfighter pilot.
(Usage) This point, where the rays pass through the perspective **plane**, is called the seat of their representation.



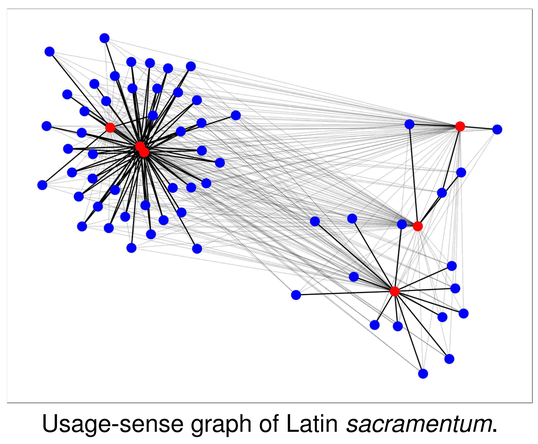
Usage-usage graph of Swedish *ledning*.



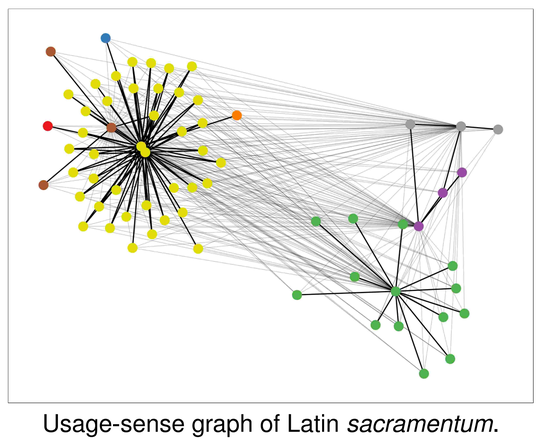
Usage-usage graph of Swedish *ledning*.

Procedure (ii): Usage-Sense Graphs

(Usage) Cum Arretinae mulieris libertatem defenderem et Cotta xviris religionem iniecisset non posse nostrum **sacramentum** iustum iudicari, [. . .]
(Sense) "a cause, a civil suit or process"

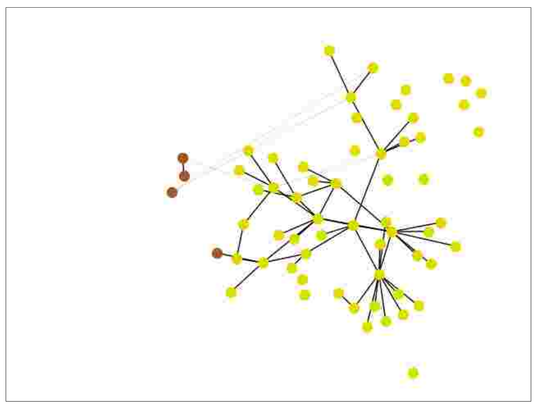


Usage-sense graph of Latin *sacramentum*.

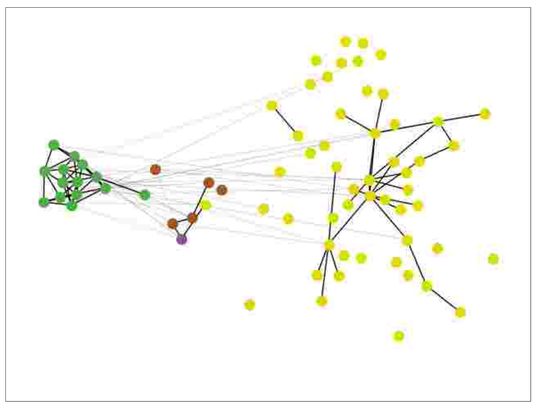


Usage-sense graph of Latin *sacramentum*.

Lexical Semantic Change



Subgraph of Swedish *ledning* for **old** subcorpus.



Subgraph of Swedish *ledning* for **new** subcorpus.

Possible Uses

- as large sets (thousands) of pairwise **semantic proximity judgments** to evaluate contextualized embeddings in multiple languages;
- the inferred change scores can be used to evaluate **semantic change detection** models;
- as **word sense disambiguation/discrimination** resources with additional aspects such as variation over time;
- graphs may be treated as research objects in their own right
- we openly release the data, clusterings, visualizations, statistics and code:

`https://www.ims.uni-stuttgart.de/data/wugs`

Acknowledgments

The authors would like to thank Diana McCarthy for her valuable input to the genesis of this task. DS was supported by the Konrad Adenauer Foundation and the CRETA center funded by the German Ministry for Education and Research (BMBF) during the conduct of this study. The creation of the data was supported by the CRETA center and the CLARIN-D grant funded by the German Ministry for Education and Research (BMBF). This task has been funded in part by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184), and *Nationella språkbanken* (the Swedish National Language Bank) – jointly funded by (2018–2024; dnr 2017-00626) and its 10 partner institutions, to NT. The Swedish list of potential change words were provided by the research group at the Department of Swedish, University of Gothenburg that work with the Contemporary Dictionary of the Swedish Academy. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. Additional thanks go to the annotators of our datasets, and an anonymous donor.

References

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89–113. doi: 10.1023/B:MACH.0000033116.57574.95
Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, *39*(3), 511–554.
Kilgarriff, A. (1998, aug). Senseval: An exercise in evaluating word sense disambiguation programs. In A. M. A. M. S. T. Thierry Fontenelle Philippe Hiligsmann (Ed.), *Proceedings of the 8th euralex international congress* (pp. 167–174). Liège, Belgium: Euralex.
McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word sense clustering and clusterability. *Computational Linguistics*, *42*(2), 245–275.
Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona, Spain: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/2020.semeval-1.1/`