



DURel Annotation Tool

Prospects on a Workbench for Lexicographers

October 19, 2022

Dominik Schlechtweg, Maïke Park

Institute for Natural Language Processing, University of Stuttgart
Leibniz-Institute for German Language, Mannheim

Motivation

- ▶ common problem:
 - ▶ given: set of word uses (corpus)
 - ▶ searched: their meanings and their relations
 - ▶ relevant for:
 - ▶ historical linguistics
 - ▶ lexicography
 - ▶ digital humanities
 - ▶ common approach: researcher scans corpus himself
 - ▶ tedious
 - ▶ subjective
 - ▶ no protocol
 - ▶ bias
 - ▶ solution: DUREL Annotation Tool¹
 - ▶ online interface
 - ▶ upload word uses for annotation
 - ▶ well-established protocol for contextual word meaning annotation
- (Erk et al., 2013; Schlechtweg et al., 2020)

¹<https://www.ims.uni-stuttgart.de/data/durel-tool>

Data

A	1824	and taking a knife from her pocket, she opened a vein in her little arm ,	😊
B	1842	And those who remained at home had been heavily taxed to pay for the arms , ammunition;	✖
C	1860	and though he saw her within reach of his arm , yet the light of her eyes seemed as far off	😊
		...	
D	1953	overlooking an arm of the sea which, at low tide, was a black and stinking mud-flat	🍷
E	1975	twelve miles of coastline lies in the southwest on the Gulf of Aqaba, an arm of the Red Sea.	🍷
F	1985	when the disembodied arm of the Statue of Liberty jets spectacularly out of the	😊

Table 1: Sample of diachronic corpus.

Annotation

- (A) [...] and taking a knife from her pocket, she opened a vein in her little **arm**, and dipping a feather in the blood, wrote something on a piece of white cloth, which was spread before her. 😊
- (D) It stood behind a high brick wall, its back windows overlooking an **arm** of the sea which, at low tide, was a black and stinking mud-flat [...]

Scale


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

Table 2: DUrel relatedness scale.

Graph representation

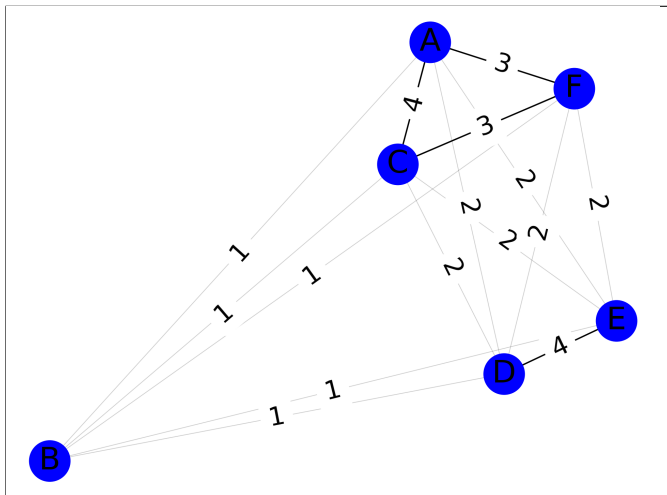


Figure 1: Word Usage Graph of English *arm*.

Clustering

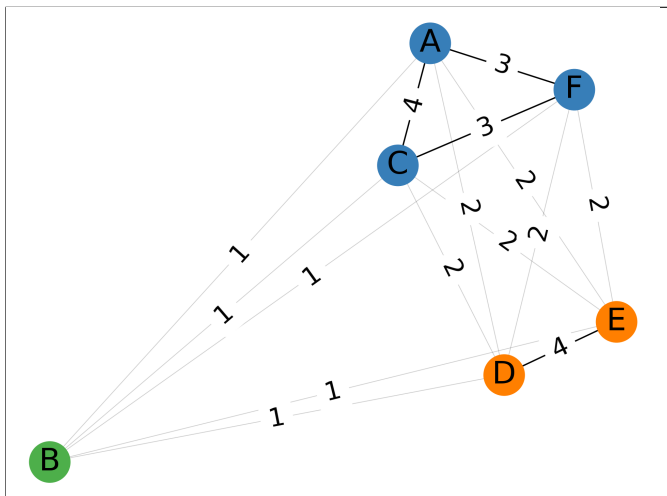
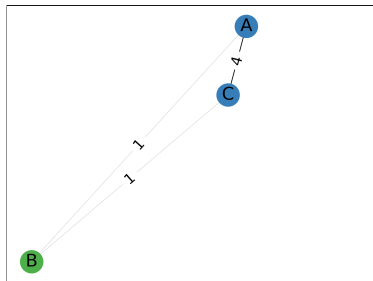
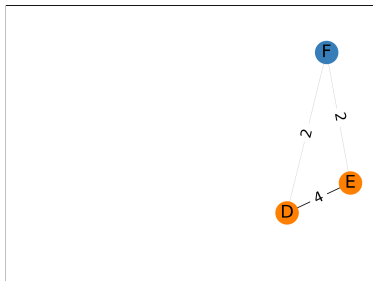


Figure 2: Word Usage Graph of English *arm*.

Lexical Semantic Change



t_1



t_2

Case Study: Lexical Semantic Change Discovery

- ▶ Kurtyigit et al. (2021)
- ▶ focus on change discovery:
 - ▶ discovering novel word senses over time in a diachronic corpus pair
 - ▶ evaluating visualizations of the annotated data from a lexicographer's point of view (how intuitive is it? are clusters conclusive? annotations reliable?)
- ▶ results:
 - ▶ high-quality predictions, high inter-annotator agreement
 - ▶ useful visualizations of clusters and relations
 - ▶ detection of previously undescribed changes that weren't included in dictionaries

Case Study: Lexical Semantic Change Discovery

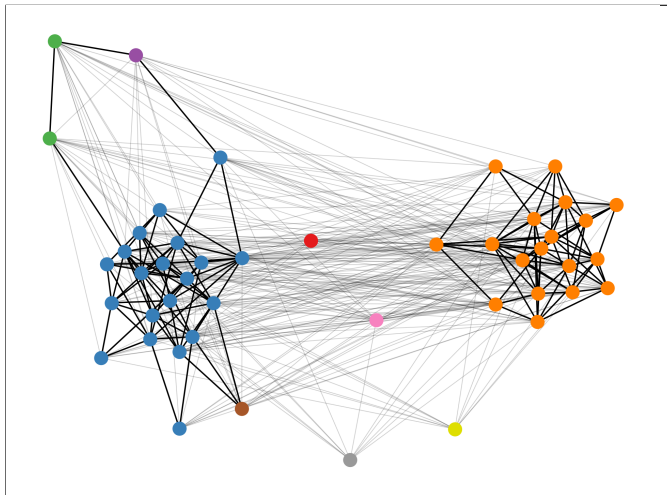
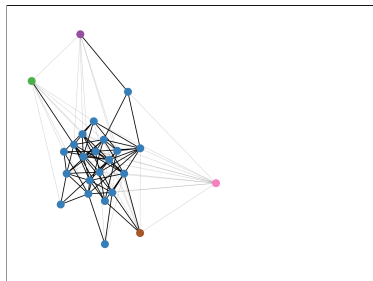
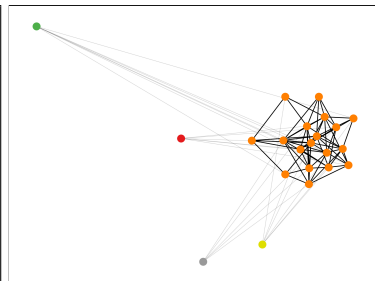


Figure 4: Word Usage Graph of German *Zehner*.

Case Study: Lexical Semantic Change Discovery



t_1



t_2

Case study: Polysemy/Synonymy across Language Varieties

- ▶ Baldissin et al. (2022)
- ▶ Diatopic lexical semantic variation in Spanish
- ▶ extend DUREl framework to onomasiological questions

Case study: Polysemy/Synonymy across Language Varieties

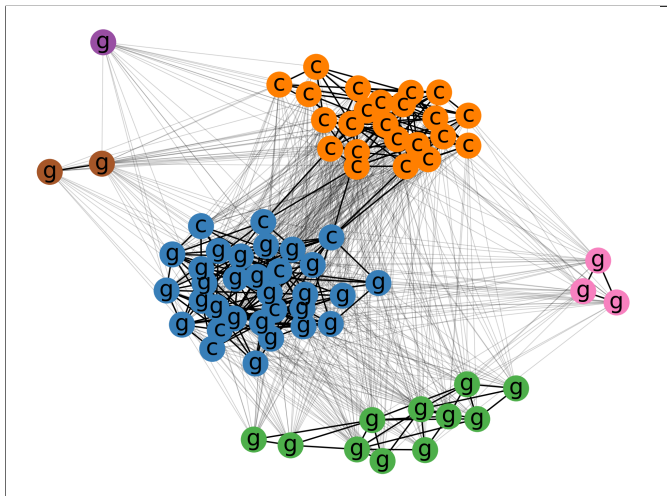
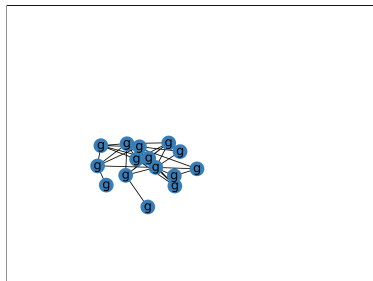
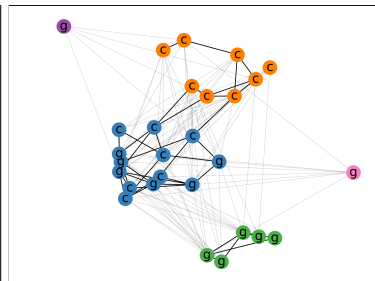


Figure 6: Word Usage Graph of Spanish *colectivo* and *guagua*.

Case study: Polysemy/Synonymy across Language Varieties



Cuba



Argentina

Further uses

- ▶ shared task on LSCD in Spanish (D. Zamora-Reina et al., 2022)
- ▶ Norwegian LSCD data set (Kutuzov et al., 2022)
- ▶ two ongoing studies from NLP and literary studies

Conclusion

- ▶ **inter-subjectivity:**
 - ▶ avoids experimenter bias through standard protocol and annotation by multiple humans
 - ▶ inter-annotator agreement gives measure of reliability
- ▶ **simple:**
 - ▶ the judgment of use pair relatedness is an intuitive task for annotators generally yielding high agreement (Erk et al., 2013; Schlechtweg et al., 2018)
 - ▶ annotated data can be visualized as semantic relatedness graphs on 2D plots
- ▶ **preparation-lean:**
 - ▶ researchers only need to sample word uses
- ▶ **grounded in theory:**
 - ▶ relatedness judgments have theoretical basis in cognitive semantics (Blank, 1997; Schlechtweg et al., 2018)
- ▶ **flexible:**
 - ▶ clustering algorithm and parameters can be changed after annotation, avoiding re-annotation

General extensions

- ▶ Optimized sampling
- ▶ Upload of data (use pairs for annotation and gold data of annotation)
- ▶ Annotation history + modifying previous annotations
- ▶ Computational Annotator
- ▶ Statistics page with annotator agreement
- ▶ Clustering methods (interface)
- ▶ Clustering visualization/analysis
- ▶ Automate prediction process for modern corpora

Extension to other tasks and fields

- ▶ traditional sense definition annotation
- ▶ use pair annotation with relation labels
- ▶ generalization of annotation scale

Lexicographic extensions

- ▶ automatic use extraction from corpora
- ▶ detection of number of senses per word
- ▶ detection of representative examples of senses
- ▶ possibility of modifying annotated data, interface
- ▶ possibility of modifying inferred sense structures (lumping/splitting of senses)
- ▶ possibility of exporting data in dictionary format (“Wörterbuch-Redaktion”); XML format is generally used;
- ▶ semantic change monitoring, large-scale annotation of random data (disagreement between users, not normally currently done, but future avenue)
- ▶ detection of multi-word units

Direct research application

- ▶ need for automation
- ▶ multiple cross-connected research lines in our project
- ▶ unique opportunity to make our research useful

Funding

- ▶ software engineer 75% for two years

References I

- Baldissin, G., Schlechtweg, D., & Schulte im Walde, S. (2022). DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish. In *Proceedings of the 13th language resources and evaluation conference*. Marseille, France: European Language Resources Association.
- Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Tübingen: Niemeyer.
- D. Zamora-Reina, F., Bravo-Marquez, F., & Schlechtweg, D. (2022). Lscdiscovery: A shared task on semantic change discovery and detection in spanish. In *Proceedings of the 3rd international workshop on computational approaches to historical language change*. Dublin, Ireland: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.lchange-1.16/>
- Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, 39(3), 511–554.
- Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021, aug). Lexical semantic change discovery. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*. Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.acl-long.543/>
- Kutuzov, A., Touileb, S., Mæhlum, P., Enstad, T. R., & Wittemann, A. (2022). *NorDiaChange: Diachronic Semantic Change Dataset for Norwegian*.
- 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/>
- Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 169–174). New Orleans, Louisiana. Retrieved from <https://www.aclweb.org/anthology/N18-2027/>