

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

# DURel Annotation Tool Measuring Patterns of Contextual Word Meaning over Time

May 14, 2021

#### Dominik Schlechtweg and Maike 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 Tool<sup>1</sup>

online interface

upload word uses for annotation

 well-established protocol for contextual word meaning annotation (Erk et al., 2013; Schlechtweg et al., 2020)

<sup>1</sup>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 Related2: 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



8

### Case study: Polysemy across Language Varieties

- Baldissin et al. (submitted)
- Diatopic lexical semantic variation in Spanish
- extend DURel framework to onomasiological questions

### Case study: Polysemy across Language Varieties



Figure 4: Word Usage Graph of Spanish *colectivo* and *guagua*.

## Case study: Polysemy across Language Varieties



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 6: Word Usage Graph of German Zehner.

## Case Study: Lexical Semantic Change Discovery



#### Upcoming case study

 target: words that recently underwent semantic change (semantic neologisms)

common problem: difficult to detect

 quantitative criteria of (semi-)automatic approaches are geared towards the lexicalization process of new words (Falk et al., 2014; Fišer & Ljubešić, 2016; Klosa & Lüngen, 2018)

tend to rely on frequency measures

common problem: difficult to define

- what does a majority of speakers perceive as new meaning (and what not)?
- which type of corpus data represents common language the most adequately?

## Targeting something really tricky to find

- target: verbs from a dictionary of neologisms in Contemporary German language
- specific problems:
  - semantic change ranges from widening/narrowing to metaphoric transfer
  - (the meaning of) a verb changes and spreads slower
  - verbs tend to be extremely infrequent in corpora used for the detection of neologisms (covering five to 10 years)
- data: newspaper texts (representing use of general German language)
- objective:
  - detect infrequent novel senses of verbs in small corpora
  - evaluate the semantic proximity between old and new senses (long-term goal: enhancing the objectivity of lexicographical decisions)

## 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

#### References I

- Baldissin, G., Schlechtweg, D., & Schulte im Walde, S. (submitted). DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish..
- Blank, A. (1997). Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen. Tübingen: Niemeyer.
- Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. Computational Linguistics, 39(3), 511–554.
- Falk, I., Bernhard, D., & Gérard, C. (2014). From non word to new word: Automatically identifying neologisms in French newspapers. In Proceedings of the ninth international conference on language resources and evaluation (LREC'14) (pp. 4337–4344). Reykjavik, Iceland: European Language Resources Association (ELRA).
- Fišer, D., & Ljubešić, N. (2016). Detecting Semantic Shifts in Slovene Twitterese. In Proceedings of recent advances in slavonic natural language processing (pp. 43–50). Czech Republic: Tribun EU.
- Klosa, A., & Lüngen, H. (2018). New German Words: Detection and Description. In Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts 17–21 (pp. 559–569). Ljubliana: Ljubljana University Press.
- Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021). Lexical Semantic Change Discovery. In Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online: Association for Computational Linguistics.
- 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/