Comparing Annotation Frameworks for Lexical Semantic Change

November 8, 2018

Dominik Schlechtweg, Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart, Germany

Motivation

- **evaluation** in research on Lexical Semantic Change Detection (LSCD) is still an unsolved issue (e.g. Cook, Lau, McCarthy, & Baldwin, 2014; Frermann & Lapata, 2016; Lau, Cook, McCarthy, Newman, & Baldwin, 2012; Takamura, Nagata, & Kawasaki, 2017)
- across languages there is no standard test set that goes beyond a few hand-selected examples
- as a result, computational models of semantic change are evaluated only superficially, while some of their predictions can be shown to be biased (Dubossarsky, Weinshall, & Grossman, 2017).

→ **we need an evaluation task definition and evaluation data**

General Criteria for Annotation

- allow calculation of agreement between annotators
- rely on clearly defined linguistic concepts
- preferably doable as a non-expert
- scale easily

Lexical Semantic Change

- LSC is inherently related to loss or emergence of word senses, as it is either:
  - **innovative**: emergence of a full-fledged additional meaning of a word, or
  - **reductive**: loss of a full-fledged meaning of a word (cf. Blank, 1997, p. 113)

$\rightarrow$ need to distinguish word senses

$\rightarrow$ problem of definition and dichotomy of word senses

Annotating LSC

- we developed **DURel** (Schlechtweg, Schulte im Walde, & Eckmann, 2018)
- yields high inter-annotator agreement of non-experts
- relies on intuitive linguistic concept of **semantic relatedness**
- it is well-grounded in cognitive semantic theory
- avoids assignment of particular sense to a word use

$\rightarrow$ requires only minimal preparation efforts

Example of Innovative Meaning Change

EARLIER	LATER
(1) *An schrecklichen Donnerwettern und heftigen Regengüssen fehlt es hier auch nicht.* 'There is no lack of horrible thunderstorms and heavy rainstorms.'	(2) a) *Oder es überschauerte ihn wie ein Donnerwetter mit Platzregen.* 'Or he was doused like a thunderstorm with a heavy shower.' b) *Potz Donnerwetter!* 'Man alive!"

Main Idea

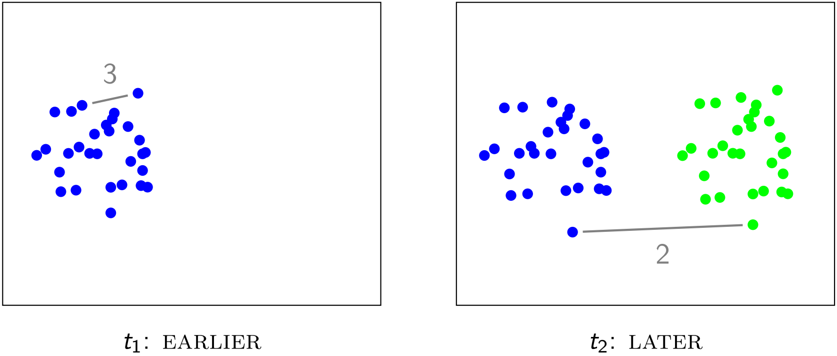


Figure 1: 2-dimensional use spaces (semantic constellation) in two time periods with a target word w undergoing innovative meaning change. Dots represent uses of w. Spatial proximity of two uses means high relatedness.

Scale

4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated
0: Cannot decide

Table 1: Four-point scale of relatedness (Schlechtweg et al., 2018).

Study details

- **five annotators** rated 1,320 German use pairs on relatedness scale in Table 1
- for **22 target words** we randomly sampled 20 use pairs per group from DTA corpus
- there are **three groups**: EARLIER (1750-1800), LATER (1850-1900) and COMPARE
- order within pairs was randomized, pairs from all groups were mixed and randomly ordered

Judgment Frequencies in Annotation Groups

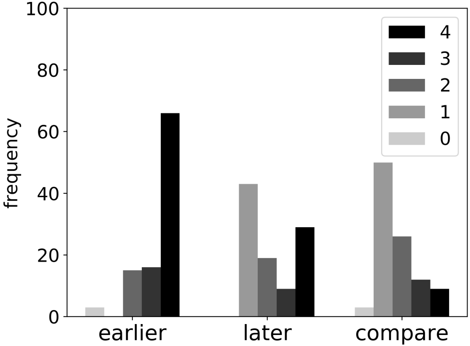


Figure 2: Judgment frequency for *Donnerwetter* (innovative).

Results – Inter-Annotator Agreement

- average pairwise correlation of **0.66**
- higher than in Erk, McCarthy, and Gaylord (2013) (between 0.55 and 0.62)

Shortcomings

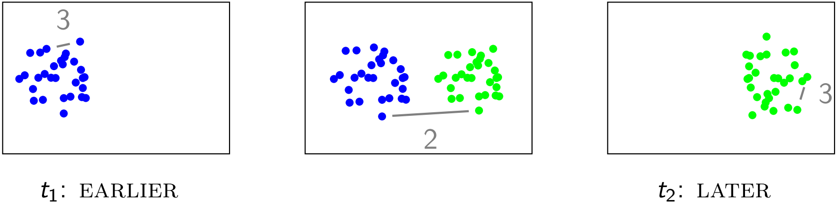


Figure 3: Innovative followed by reductive meaning change. Mean relatedness change predicts no LSC.

Alternative Annotation Strategy

- the above-examined measure of change collapses in certain semantic constellations
- how can we improve this?

$\rightarrow$ we will try to retrieve the underlying sense frequency distributions

Choosing a Target Word and Time Periods

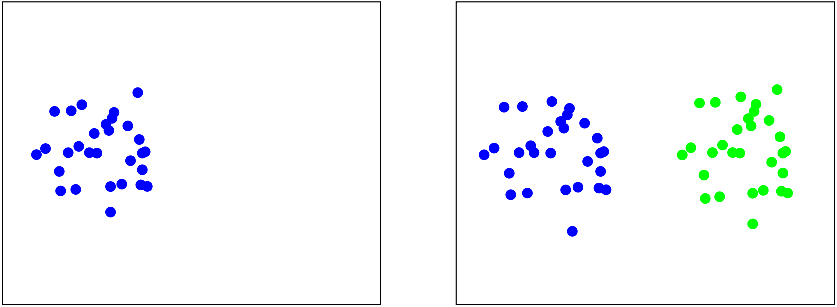
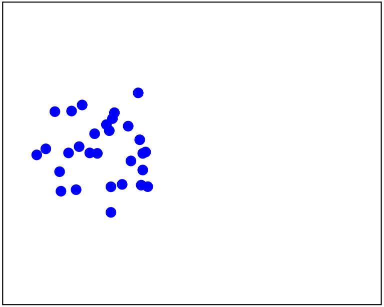
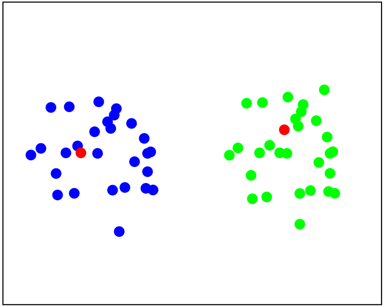

t_1: EARLIER $\qquad$ t_2: LATER

Figure 4: Underlying semantic constellation for a target word.

Choosing Centroids



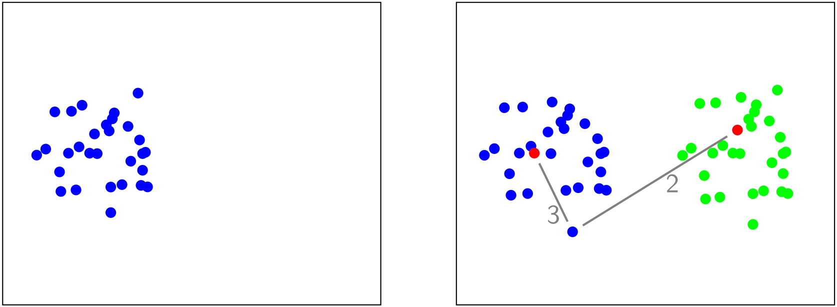
t_1: EARLIER



t_2: LATER

Figure 5: Sense centroids for each sense cluster.

Comparing Uses

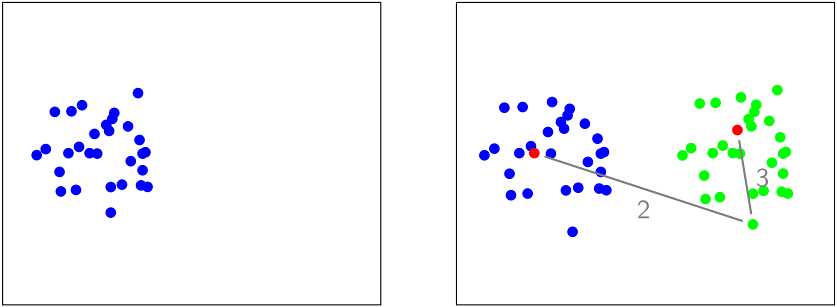


t_1: EARLIER

t_2: LATER

Figure 6: Comparison of uses from different time periods against sense centroids.

Comparing Uses



t_1: EARLIER $\qquad$ t_2: LATER

Figure 7: Comparison of uses from different time periods against sense centroids.

Comparing Uses

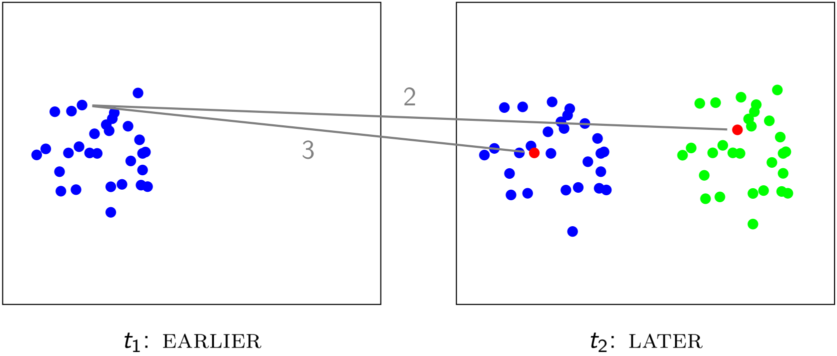


Figure 8: Comparison of uses from different time periods against sense centroids.

Pros and Cons

- advantages:
  - still graded assignment
  - centroids represent meanings (no definition needed)
  - centroids can be chosen to be clear and distinguishable contexts
  - graphs provide an accessible method of visualization
- disadvantages:
  - strong assumption of clear-cut clusters and good choice of centroids
  - annotation time increases sharply with polysemous words
- annotation is test for whether uses can be assigned to different clusters as represented by the chosen centroids
- annotators either verify or falsify the choice of centroids
- bad choices will be obvious from the annotated data

Study Details

- the annotation is carried out maximally parallel to Schlechtweg et al. (2018) (i.e., same guidelines, scale, annotators, target lemmas, time periods)
- the only difference is **sampling process**:
  1. choose a target lemma and time periods
  2. sample 10 contexts for each time period (EARLIER and LATER)
  3. choose centroid uses
  4. combine each use with each centroid into a use pair
  5. combine each centroid with each other centroid
  6. switch the order of every second pair and randomly shuffle all pairs
- by this we obtain a total of 788 use pairs

Retrieval of Sense Frequency Distributions

- can be tricky in the case of e.g. equivocal judgments
- sources of conflict are
  - uses assigned to more than one centroid,
  - uses assigned to none of the centroids,
  - centroids judged not to be clearly distinct,
  - zero-judgments (incomprehensible),

$\rightarrow$ we need a way to deal with these cases

Retrieval of Sense Frequency Distributions

- we deal with these cases in the following way:
  1. zero-judgments are ignored,
  2. if there are centroid pairs with mean judgments $>= 2.5$, they are treated as representing the same meaning,
  3. centroids are collapsed transitively,
  4. uses with a mean judgment with a certain centroid $>= 2.5$ will be assigned to that centroid,
  5. if a use is assigned to more than one centroid, the one with the highest judgment is chosen,
  6. if a use is assigned to none of the centroids, it is treated as representing an additional meaning

Retrieval of Word Sense Distributions

- with this algorithm we can automatically retrieve sense frequency distributions from the annotated data
- if the data doesn't allow to do this safely, the algorithm will provide us with the necessary knowledge to exclude the data/revise the annotation style
- the data can be conveniently visualized as (spatial plots of) usage graphs constructed by the annotation data
- the inferred sense frequency distributions show up as distinct clusters of uses in the spatial plots of the respective usage graphs

Annotation Results – Some Examples

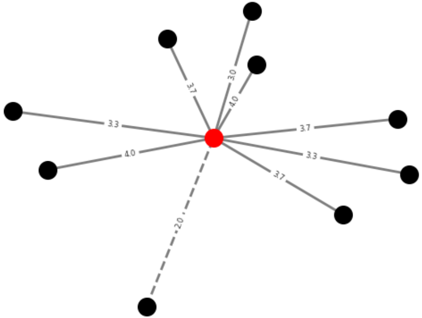


Figure 9: Graph visualization retrieved from annotation data from EARLIER time period for target *Abend*. Centroids are plotted red. Continuous lines mark edge judgments $>=2.5$, while dashed lines mark edge weights $<=2.5$. Node distance between connected! nodes (mostly) reflects their judgment score (edge label).

Examples

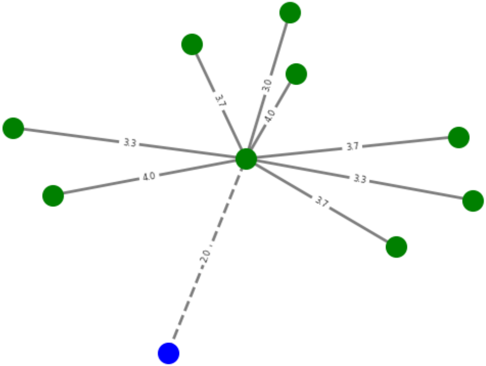


Figure 10: Graph visualization of EARLIER time period for target *Abend* with inferred distribution: $T_1 = (1, 9)$. Different colors mark uses of different meanings.

Examples

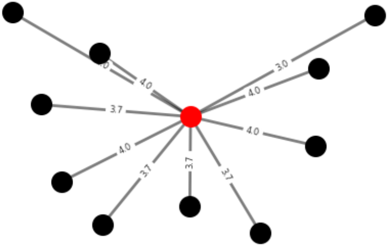


Figure 11: Graph visualization of LATER time period for target *Abend*.

Examples

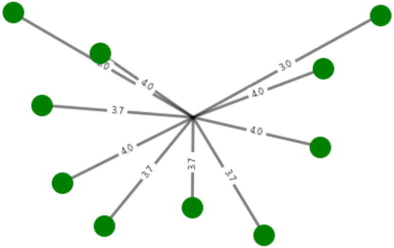


Figure 12: Graph visualization of LATER time period for target *Abend* with inferred distribution: $T_2 = (0, 10)$.

Examples

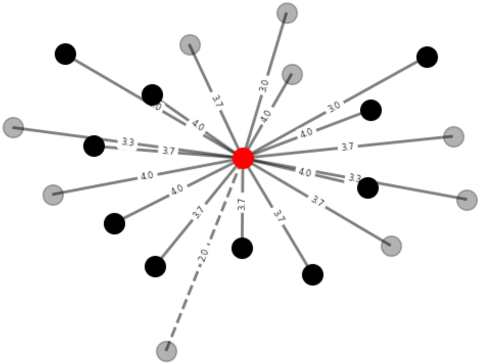


Figure 13: Graph visualization of LATER time period for target *Abend*. Inferred distributions $T_1 = (1, 9)$ and $T_2 = (0, 10)$. Transparent nodes mark uses from t_1 (EARLIER).

Examples

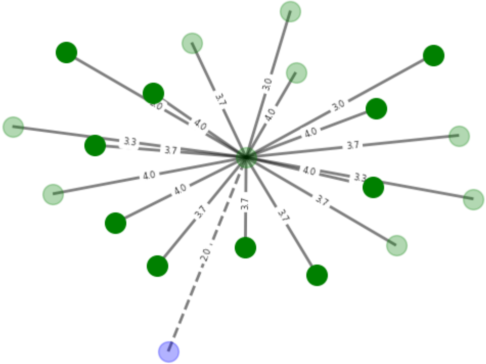


Figure 14: Graph visualization of LATER time period for target *Abend*. Inferred distributions $T_1 = (1, 9)$ and $T_2 = (0, 10)$.

Examples

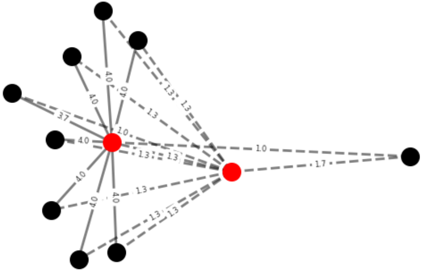


Figure 15: Target: *Vorwort*. Time period: t_1.

Examples

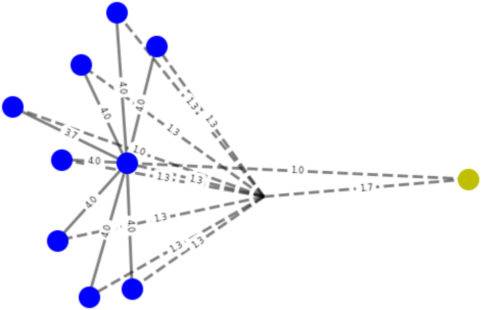


Figure 16: Target: *Vorwort*. Time period: t_1. Distribution: $T_1 = (9, 0, 1)$

Examples

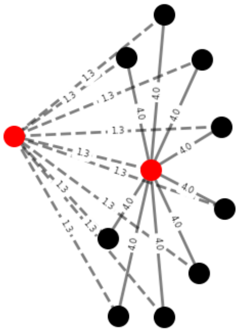


Figure 17: Target: *Vorwort*. Time period: t_2.

Examples

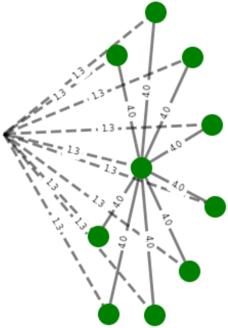


Figure 18: Target: *Vorwort*. Time period: t_2. Distribution: $T_2 = (0, 10, 0)$

Examples

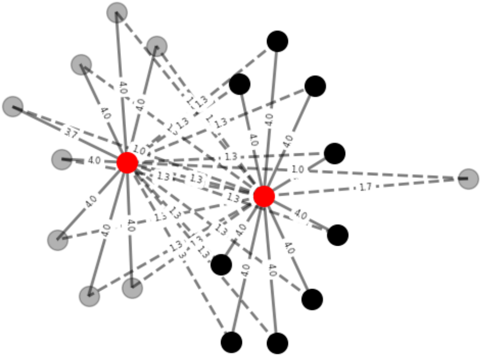


Figure 19: Target: *Vorwort*.

Examples

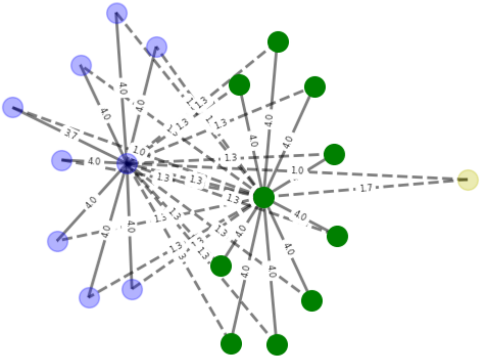


Figure 20: Target: *Vorwort*. Inferred distributions $T_1 = (9, 0, 1)$ and $T_2 = (0, 10, 0)$.

Examples

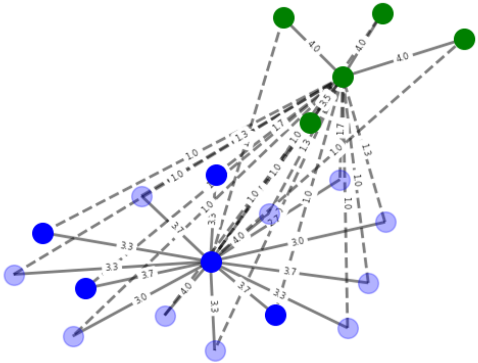


Figure 21: Target: *billig*. Inferred distributions $T_1 = (10, 0)$ and $T_2 = (5, 5)$.

Examples

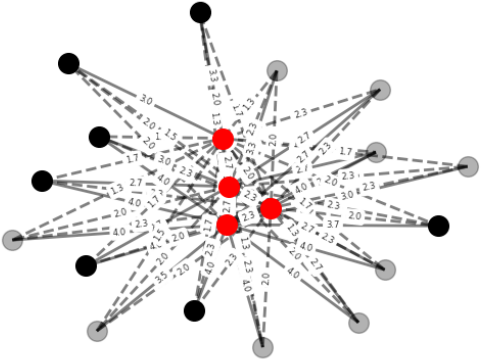


Figure 22: Target: *geharnischt*.

Examples

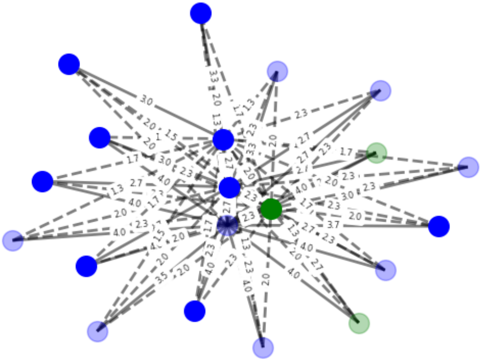


Figure 23: Target: *geharnischt*. Inferred distributions $T_1 = (8, 2)$ and $T_2 = (9, 1)$.

Annotation Results – Inter-Annotator Agreement

- average pairwise correlation of **0.72**
- higher than in Schlechtweg et al. (2018) (0.66)

Overview

	across all targets
centroids collapsed	8/43 (14/22 targets with > 1 centroids)
centroid conflicts	2
use conflicts	37/397
uses excluded due to 0-judgment	17/397
uses finally uniquely assigned	363/380
uses finally multiply assigned	17/380
assigned by maximum judgment	13/17
randomly assigned	4/17

Table 2: Overview of annotation results with conflicts.

Some Conclusions

- it generally works
- data can be iteratively revised
- centroids should be checked iteratively with annotators before starting the annotation
- **if you want to work with DURel, please write me an email!**

Bibliography

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Tübingen: Niemeyer.

Cook, P., Lau, J. H., McCarthy, D., & Baldwin, T. (2014). Novel word-sense identification. In *25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland* (pp. 1624–1635).

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1147–1156). Copenhagen, Denmark.

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring word meaning in context. *Computational Linguistics*, *39*(3), 511–554.

Frermann, L., & Lapata, M. (2016). A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, *4*, 31–45.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., & Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 591–601). Stroudsburg, PA, USA.

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.* New Orleans, Louisiana.

Takamura, H., Nagata, R., & Kawasaki, Y. (2017). Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (pp. 1195–1204). Valencia, Spain: Association for Computational Linguistics.