Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change

Dominik Schlechtweg, *[*]* **Sabine Schulte im Walde,** *[†]* **Stefanie Eckmann**
*[*]*Institute for Natural Language Processing, University of Stuttgart, Germany
*[†]*Historical and Indo-European Linguistics, LMU Munich, Germany
dominik.schlechtweg@ims.uni-stuttgart.de, schulte@ims.uni-stuttgart.de,
stefanie.eckmann@campus.lmu.de

Abstract

We propose a framework that extends synchronic polysemy annotation to diachronic changes in lexical meaning, to counteract the lack of resources for evaluating computational models of lexical semantic change. Our framework exploits an intuitive notion of semantic relatedness, and distinguishes between innovative and reductive meaning changes with high inter-annotator agreement. The resulting test set for German comprises ratings from five annotators for the relatedness of 1,320 use pairs across 22 target words.

1 Introduction

We see an increasing interest in the automatic detection of semantic change in computational linguistics (Hamilton et al., 2016; Frermann and Lapata, 2016; Schlechtweg et al., 2017, i.a.), motivated by expected performance improvements of practical NLP applications, or theoretical interest in language or cultural change. However, a major obstacle in the computational modeling of semantic change is evaluation (Lau et al., 2012; Cook et al., 2014; Frermann and Lapata, 2016; Dubossarsky et al., 2017). Most importantly, there is no reliable test set of semantic change for any language that goes beyond a small set of hand-selected targets. We counteract this lack of resources by extending a framework of synchronic polysemy annotation to the annotation of Diachronic Usage Relatedness (DURel). DURel has a strong theoretical basis and at the same time makes use of established synchronic procedures that rely on the intuitive notion of semantic relatedness. The annotations distinguish between innovative and reductive meaning change with high inter-annotator agreement. DURel is language-independent and thus applicable across languages; this paper introduces the first test set of lexical semantic change for German.

2 Related Work

A large number of studies has been performed on synchronic word sense annotation (see Ide and Pustejovsky, 2017 for an overview). Within this set, our paper is most related to work focusing on graded polysemy annotation. Most prominently, Soares da Silva (1992) is interested in the question of whether the theoretical distinction between polysemy and homonymy can be experimentally verified; Brown (2008) wants to know how fine-grained word senses are, and Erk et al. (2009, 2013) examine whether we should adopt a graded notion of word meaning.

In contrast, there is little work on annotation with a focus on semantic change, despite the growing interest and modeling efforts in the field of semantic change detection. Lau et al. (2012) and Cook et al. (2014) aim at verifying the semantic developments of their targets by a quasi-annotation procedure of dictionary entries, however without reporting inter-annotator agreement or other measures of reliability. Gulordava and Baroni (2011) ask annotators for their intuitions about changes but without direct relation to language data. Bamman and Crane (2011) exploit aligned translated texts as source of word senses and conduct a very limited annotation study on Latin texts from different time periods. Schlechtweg et al. (2017) propose a small-scale annotation of metaphoric change, but altogether there is no standard test set across languages that goes beyond a few hand-selected examples.

3 Lexical Semantic Change

It is well-known that lexical semantic change and polysemy are tightly connected. For example, Blank (1997) develops an elaborate theory where polysemy is the synchronic, observable result of

Proceedings of NAACL-HLT 2018, pages 169–174
New Orleans, Louisiana, June 1 - 6, 2018.

lexical semantic change. He distinguishes two main types of lexical semantic change:

- **innovative meaning change**: emergence of a full-fledged additional meaning of a word; old and new meaning are related by polysemy
- **reductive meaning change**: loss of a full-fledged meaning of a word

An example of innovative meaning change is the emergence of polysemy in the German word *Donnerwetter* around 1800 (Paul, 2002). Before ≈1800 *Donnerwetter* was only used in the meaning of 'thunderstorm'. After 1800 we still observe this meaning, and in addition we find a new, clearly distinguished meaning as a swear word 'Man alive!'. An example of reductive meaning change is the German word *Zufall*. It had two meanings ≈1850, 'seizure' and 'coincidence' (Osman, 1971). After 1850, the word occurs less and less often in the former meaning, until it is exclusively used in the meaning of 'coincidence'. *Zufall* lost the meaning 'seizure'.

3.1 Semantic Proximity

Based on Prototype Theory (Rosch and Mervis, 1975), Blank develops criteria to decide whether word uses are related by polysemy. He defines a continuum of *semantic proximity* with polysemy located between identity and homonymy, as depicted in Table 1.

↑
Identity
Context Variance
Polysemy
Homonymy

Table 1: Continuum of semantic proximity (cf. Blank, 1997, p. 418).

While it is difficult to directly apply these criteria to practical annotation tasks, we exploit the scale of semantic proximity indirectly, as previously done by synchronic research on polysemy applying similar scales (Soares da Silva, 1992; Brown, 2008; Erk et al., 2013). Especially Erk et al.'s in-depth study validates an annotation framework relying on a scale of semantic proximity, revealing high inter-annotator agreement and strong correlation with traditional single-sense annotation as well as annotation of multiple lexical paraphrases. For our study, we decided to adopt a relatedness scale similar to Brown's, shown in Table 2.

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

Table 2: Our 4-point scale of relatedness derived from Brown (2008).

3.2 Diachronic Usage Relatedness (DURel)

We frame our interest in lexical semantic change as judging the strength of semantic relatedness across use pairs of a target word w within a specific time period t_i. A high mean proximity value indicates meaning identity or context variance, while a low value indicates polysemy or homonymy, cf. Table 1. This strategy is applied independently to two time periods t_1 and t_2, as illustrated in Figure 1. Innovative vs. reductive meaning change can then be measured by decrease vs. increase in the mean relatedness value of w from t_1 to t_2. To see why this is justified, consider the different semantic constellations of w's use pairs in t_1 and t_2 in Figure 1. If w is monosemous in t_1 and undergoes innovative meaning change between the two time periods, we expect to find use pairs in the later period t_2 combining the old and new meaning which are less related (score: 2) than the use pairs from the earlier period t_1 (score: 3). According to this rationale, the mean relatedness values across w's use pairs should be lower in t_2 than in t_1. The reverse applies to reductive meaning change.

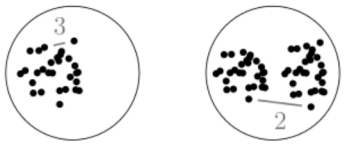


Figure 1: Two-dimensional use spaces (Tuggy, 1993; Zlatev, 2003) in two time periods with a target word w undergoing innovative meaning change. Dots represent uses of w. Spatial proximity of two uses means high relatedness.

There are a number of other, more complex se-

mantic constellations. For example, if w not only gains a new meaning, but rapidly loses the old meaning, we cannot necessarily expect the mean relatedness of w's use pairs to be higher in the later than in the earlier time period. In order to cover such cases, we will not only measure the mean relatedness within the EARLIER and the LATER groups of use pairs but also in a mixed COMPARE group where each pair consists of a use from the EARLIER and a use from the LATER group. By this, old and new meaning are directly compared, and we do not have to rely on the assumption that the old meaning is still present.

By applying the above-described procedure to all target words and sorting them according to their mean relatedness scores, we obtain a ranked list for each of the three groups EARLIER, LATER and COMPARE. We then exploit two measures of change: (i) ΔLATER measures changes in the degree of mean relatedness of words, and is derived by subtracting a target w's mean in EARLIER from its mean in LATER: $\Delta\text{LATER}(w) = Mean_l(w) - Mean_e(w)$. Positive vs. negative values on this measure indicate innovative vs. reductive meaning change. (ii) COMPARE directly measures the relatedness *between* the EARLIER and the LATER group and thus corresponds to w's mean in the COMPARE group: $\text{COMPARE}(w) = Mean_c(w)$. High vs. low values on COMPARE indicate weak vs. strong change, where the change includes both innovative and reductive meaning changes.

4 Annotation Study

Five native speakers of German were asked to rate 1,320 use pairs on our 4-point scale of relatedness in Table 2. All annotators were students of linguistics. We explicitly chose two annotators with a background in historical linguistics in order to see whether knowledge about historical linguistics has an effect on the annotation. Annotators were not told that the study is related to semantic change.[1]

Target Words. The target words were selected by manually checking a corpus for innovative and reductive meaning changes, based on cases of metaphoric, metonymic change and narrowing (innovative) as reported by Paul (2002), and cases of reduction due to homonymy (reductive) as reported by Osman (1971). The corpus we used is DTA (Deutsches Textarchiv, 2017), a freely available diachronic corpus of German. By focusing on a late time period (19th century), we tried to reduce problems coming with historical language data as much as possible. We still normalized special characters to modern orthography.

	target sentence 1	...	target sentence 2
1	Bemerkungen. Ein **Donnerwetter** in Paris ist mit so vielen Verdrieslichkeiten verknüpft, daß ichs hier anführen muß. Wir hatten heute Abends eins von 6. Uhr bis halb 11. Uhr des Nachts.		Der andre observirte schärfer mit dem Ausruf: „**Donnerwetter**, sollte ich mich irren! Sie changirt nicht Farbe, und doch zuckte sie zusammen, als die Lupinus ihr was ins Ohr sagte."

Figure 2: Use pair from annotation table (English adaption).

We included only those words as targets for which we found the change suggested by the literature reflected in the corpus, either weakly or strongly, because an annotation relying on a random selection of words suggested to undergo change is likely to produce a set with very similar and rather low values representing small effects. We thus guaranteed to include both: words for which we expected weak effects as well as words for which we expected strong effects. We ended up with 19 cases of innovation and 9 cases of reduction. Three words, *Anstalt*, *Anstellung* and *Vorwort* represent especially interesting cases and were selected more than once for the test set since they undergo both innovative *and* reductive change between the investigated time periods.

Sampling. For each target word we randomly sampled 20 use pairs from DTA (searching for the respective lemma and POS) for each of the groups EARLIER (1750-1800), LATER (1850-1900) and COMPARE, yielding 60 use pairs per word and 1,320 use pairs for 22 target words in total.

A use of a word is defined as the sentence the word occurs in. The annotators were provided these sentences as well as the preceding and the following sentence in the corpus, cf. Figure 2. We double-checked that each use of a word was only sampled once within each group. If the total number of uses in the group was less than needed, uses were allowed twice across pairs. Before presenting the use pairs to the annotators in a spreadsheet, uses within pairs were randomized, and pairs from all groups were mixed and randomly ordered.

[1] The guidelines (adapted from the synchronic study by Erk et al., 2013) and the experiment data are publicly available at www.ims.uni-stuttgart.de/data/durel/.

5 Results

Agreement. In line with Erk et al. (2013) we measure inter-annotator agreement as the average over pairwise Spearman's ρ correlations (omitting 0-judgments), cf. Table 3. The bottom line provides the agreement of each annotator's judgments against the average judgment score across the other annotators. The range of correlation coefficients is between 0.57 and 0.68, with an average correlation of 0.66. All the pairs are highly significantly correlated ($p < 0.01$).

	1	2	3	4	5
1		0.59	0.63	0.67	0.66
2			0.57	0.64	0.65
3				0.64	0.62
4					0.68
avg	0.71	0.68	0.68	0.75	0.74

Table 3: Correlation matrix for pairwise correlation agreement; *avg* refers to agreement of the annotator in the column against the average across the other annotators.

The annotators with historical background are annotators 4 and 5, who show the highest pairwise agreement and also the highest agreement with the average of the other annotators. This indicates that historical knowledge makes a positive difference when annotating DURel. Yet, the agreements of the non-expert annotators only deviate slightly.

Overall, our correlations are comparable and even moderately higher than the ones found in Erk et al. (2013), who report average correlation scores between 0.55 and 0.62. This difference is remarkable, given that annotators had to judge historical data. Note, however, that the studies are not exactly comparable, as Erk et al. used a more fine-grained 5-point scale, and we presumably excluded a larger number of 0-judgments.

Qualitative Analysis. Figure 3 shows the target words ranked according to their values on ΔLATER. We can clearly identify three groups: words with values >0, <0, and a majority with values ≈ 0 difference in mean between the earlier and the later time period. The three topmost words have previously been classified as reductive, the three lowermost words as innovative meaning changes.

Figure 4 compares the distributions across relatedness scores for our two example words *Zufall* and *Donnerwetter* from above. In EARLIER, *Zufall*'s ratings (i.e., the number of times a specific rating 0–4 was provided) vary much more than in LATER where it has a high number of 4-judgments. The contrary is the case for *Donnerwetter*. In addition, we find a clear difference between the two words in the COMPARE group, because *Donnerwetter* is used in a variety of new figurative ways in LATER, while *Zufall*, besides losing the meaning 'seizure', retains the prevalent meaning 'coincidence' in both time periods.

Upon closer inspection, the words deviating most from our predictions show either that the change is already present before 1800 (e.g., *Steckenpferd*, 'toy > toy; hobby'), that the new meaning has a very low prevalence (e.g., *Museum*, 'study room; arts collection > arts collection'), or that there are additional, previously not identified uses in the later time period (e.g., *Feine*, 'fineness; grandeur > grandeur'). The mean value for reduction is 0.39, while it is -0.18 for innovation.

Overall, these findings confirm our predictions and validate ΔLATER as a measure of lexical semantic change. The case of *Presse*, 'printing press > printing press; print product/producer', however, shows its shortcomings: ΔLATER wrongly predicts no change for *Presse*, although it is clearly present, because the new meaning has a very high prevalence. ΔLATER cannot capture such cases, while COMPARE can: it predicts strong change for *Presse*.

Since COMPARE measures the degree of change rather than distinguishing between types of change, the highest values in its ranked list refer to cases with values$\approx$0 in ΔLATER, and the lowest values refer to cases with extreme values of ΔLATER. A special case is *Feder* 'bird feather > bird feather; steel clip', which reveals the need for normalization of the COMPARE-measure: the word is highly polysemous and has approximately the same distribution in every group, because the new meaning 'steel clip' has a very low prevalence. For ΔLATER this correctly leads to a 0-prediction. In contrast, COMPARE predicts strong change, because due to polysemy there is a high probability to sample distantly related use pairs in the COMPARE group.

Discussion. While our measures enable us to predict various semantic change constellations, we also demonstrated that they collapse in cer-

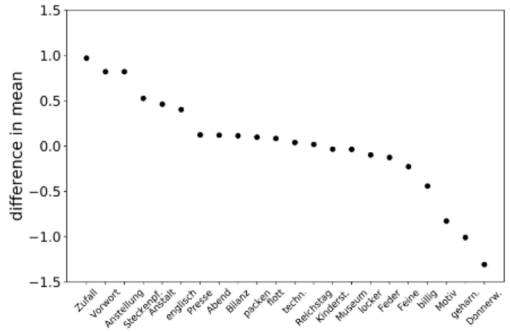
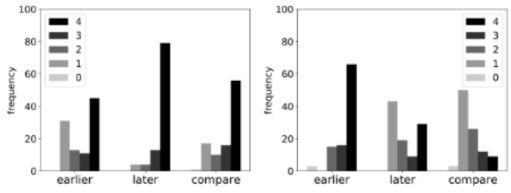


Figure 3: ΔLATER: Rank of target words according to increase in mean usage relatedness from EARLIER to LATER.



Zufall (reductive) *Donnerwetter* (innovative)

Figure 4: Plots of judgment freq. for target words per group.

tain semantic constellations: ΔLATER is accurate when used for simple semantic constellations (i.e., only one reductive or innovative meaning change), where the old meaning roughly maintains its prevalence, thus making ΔLATER be prone to corpus effects such as changes in text genre. For an optimal application of this measure we therefore recommend (i) to choose directly adjacent and short time periods for annotation, as the number of changes increases with the length of the time period, and (ii) to use a well-balanced corpus for the annotation, ideally across all periods.

Unlike ΔLATER, COMPARE has the advantage to capture multiple changes over time, but it confuses polysemy and meaning change. In future work, we aim to solve this issue by normalizing COMPARE with a measure of polysemy: For any target word w the values from the EARLIER group determine its degree of polysemy in the earlier time period. Hence, the normalized $\Delta\text{COMPARE}(w) = Mean_c(w) - Mean_e(w)$ intuitively measures how much the values in the COMPARE group differ from what we would already expect from w's early polysemy, so it predicts no change in the case of a stable polysemous word, and it predicts change if the word gains or loses a meaning.

6 Conclusion

This paper presented a general framework DURel for language-independent annotation of Diachronic Usage Relatedness, in order to develop test sets for lexical semantic change. In addition to a strong theoretical basis, DURel shows empirical validity in our annotation study with high inter-annotator agreement. It relies on an intuitive notion of semantic relatedness and needs no definition of word senses.

Furthermore, we proposed two measures of lexical semantic change that predict various semantic change constellations. While one measure successfully distinguishes between innovative and reductive meaning change, we also demonstrated the need to refine and normalize the measures in order to capture more variants of constellations regarding the interplay of polysemy and meaning reduction/innovation.

The annotated test set for German is publicly available and can be used to compare computational models of semantic change, and more generally to evaluate models of lexical variation in corpora across times, domains, etc. Further test sets across languages can be obtained by applying DURel to the respective language uses.

Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732. We thank Katrin Erk, Diana McCarthy and Susan Brown Windisch for providing their expertise, experience and data; Eleonore Brandner, Jun Chen, Fabian Bross and Diego Frassinelli for helpful discussions and comments; Delia Alf, Pavlos Musenidis, Cornelia van Scherpenberg and Jennifer Schild for help with the annotations; and the three reviewers for their constructive criticism.

References

D. Bamman and G. Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. ACM, New York, NY, USA, pages 1–10.

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

S. W. Brown. 2008. Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Stroudsburg, PA, USA, pages 249–252.

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

Deutsches Textarchiv. 2017. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften. http://www.deutschestextarchiv.de/.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 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*. Copenhagen, Denmark, pages 1147–1156.

K. Erk, D. McCarthy, and N. Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. Stroudsburg, PA, USA, pages 10–18.

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

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

K. Gulordava and M. Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Stroudsburg, PA, USA, pages 67–71.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, pages 2116–2121.

N. Ide and J. Pustejovsky, editors. 2017. *Handbook of Linguistic Annotation*. Springer, Dordrecht.

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

N. Osman. 1971. *Kleines Lexikon untergegangener Wörter: Wortuntergang seit dem Ende des 18. Jahrhunderts*. Beck, München.

H. Paul. 2002. *Deutsches Wörterbuch: Bedeutungsgeschichte und Aufbau unseres Wortschatzes*. Niemeyer, Tübingen, 10. edition.

E. Rosch and C.B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7:573–605.

D. Schlechtweg, S. Eckmann, E. Santus, S. Schulte im Walde, and D. Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning*. Vancouver, Canada, pages 354–367.

A. Soares da Silva. 1992. Homonímia e polissemia: Análise sémica e teoria do campoléxico. In *Actas do XIX Congreso Internacional de Lingüística e Filoloxía Románicas*. Fundación Pedro Barrié de la Maza, La Coruña, volume 2 of *Lexicoloxía e Metalexicografía*, pages 257–287.

David Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics* 4(3):273–290.

J. Zlatev. 2003. *Polysemy or generality? Mu*, Mouton de Gruyter, volume 23 of *Cognitive Approaches to Lexical Semantics*, pages 447–494.