



University of Stuttgart
Germany



The LSCD Benchmark: A testbed for diachronic word meaning tasks

September 7, 2023

Dominik Schlechtweg

Institute for Natural Language Processing, University of Stuttgart

Introduction

- ▶ Lexical Semantic Change Detection (Schlechtweg, 2023)
 - ▶ goal: automate the analysis of changes in word meanings over time
 - (1) *Der zweyte Theil vom Bauernrechte ist schon lange aus der **Presse**;*
'The second part of Farmers' Rights already left the **press**;'
 - (2) *Alle Freiheiten suspendirt! die persönliche Freiheit wie die der **Presse**!*
'All freedoms suspended! the personal freedom as well as the one of the **press**!'
- ▶ **heterogeneity** and **modularity** in models, datasets and tasks
- create one repository¹ standardizing model component combinations, dataset preprocessing and evaluation

¹<https://github.com/ChangeIsKey/LSCDBenchmark>

Human Measurement of Lexical Semantic Change

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 , ammuntion;	✖
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.

Word Use Pairs

- (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 [...]

Semantic Proximity Scale


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

Table 2: DUrel relatedness scale.

Graph representation

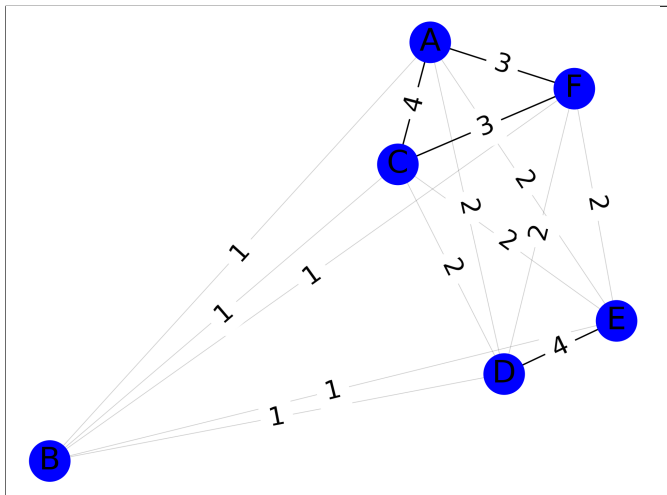


Figure 1: Word Usage Graph of English *arm*.

Clustering

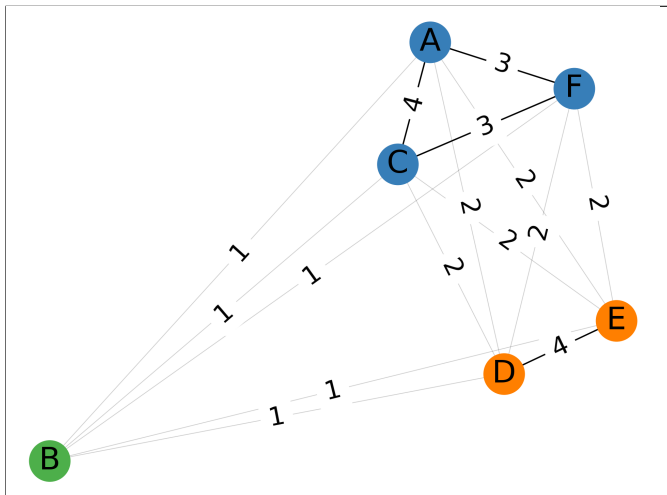
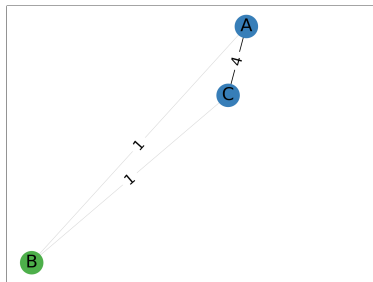
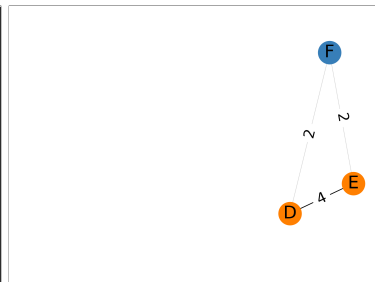


Figure 2: Word Usage Graph of English *arm*. $D = (3, 2, 1)$.

Lexical Semantic Change



$t_1, D_1 = (2, 0, 1)$



$t_2, D_2 = (1, 2, 0)$

Change Scores

- ▶ **binary change** (loss and gain of senses)
- ▶ **graded change** (changes in sense probabilities)

Example: Swedish *ledning*²

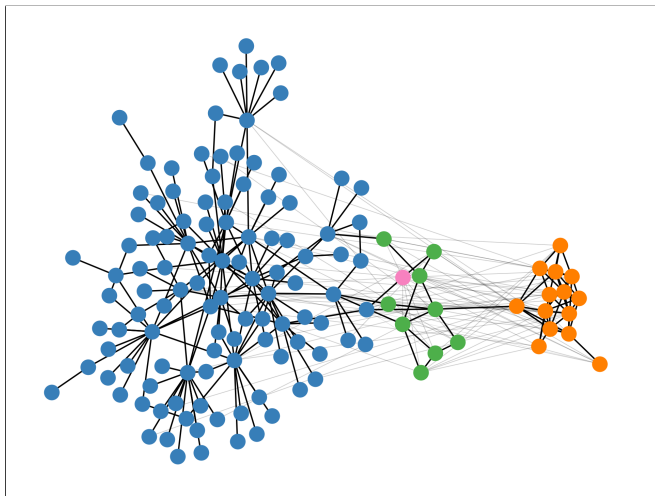


Figure 4: WUG of Swedish *ledning*.

²Datasets available at <https://www.ims.uni-stuttgart.de/data/wugs>

Example: Swedish *ledning*

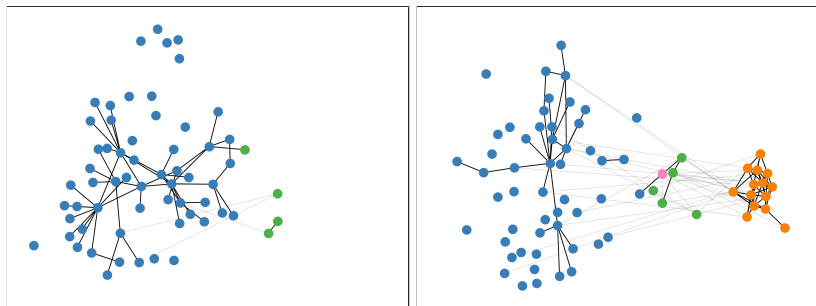


Figure 5: WUGs of Swedish *ledning*: subgraphs for 1st time period G_1 (left) and 2nd time period G_2 (right). $D_1 = (58, 0, 4, 0)$, $D_2 = (52, 14, 5, 1)$, $B(w) = 1$ and $G(w) = 0.34$.

Example: German *Eintagsfliege*

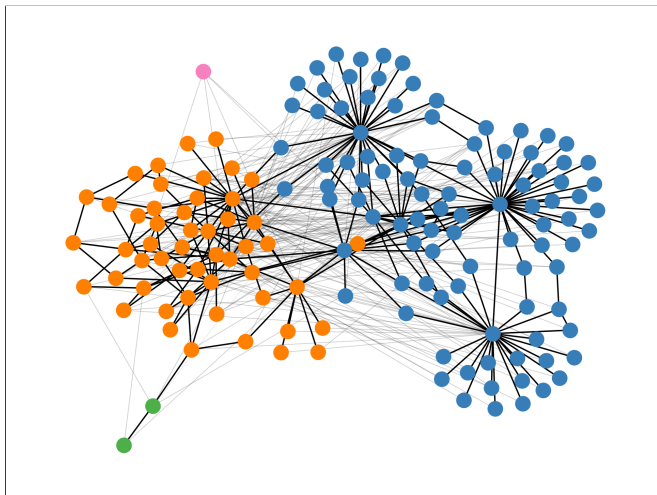


Figure 6: WUG of German *Eintagsfliege*.

Example: German *Eintagsfliege*

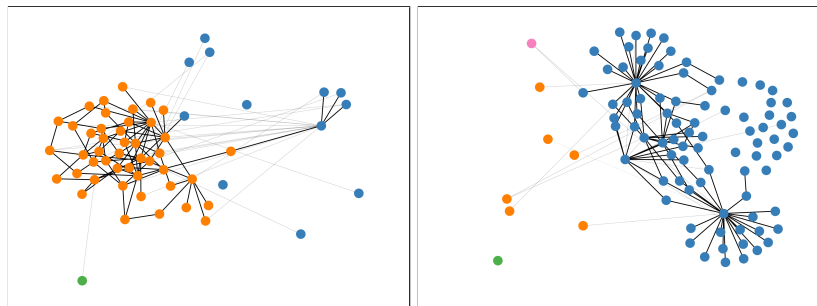


Figure 7: WUG of German *Eintagsfliege*: subgraphs for 1st time period G_1 (left) and 2nd time period G_2 (right). $D_1 = (12, 45, 0, 1)$, $D_2 = (85, 6, 1, 1)$, $B(w) = 0$ and $G(w) = 0.66$.

Summary of Annotation Steps

1. semantic proximity labeling
2. clustering
3. change measurement

Summary of Annotation Steps with Tasks

1. semantic proximity labeling ↔ **Word-in-Context Task**
2. clustering ↔ **Word Sense Induction**
3. change measurement ↔ **Lexical Semantic Change Detection** (including previous tasks)

Computational Measurement of Lexical Semantic Change

- ▶ Typical (token-based) Model is composed by
 1. semantic proximity model (e.g. similarity between contextualized embeddings)
 2. clustering method (optional)
 3. change measure

The LSCD Benchmark

- ▶ standardized implementation of
 - ▶ model components (modularity/flexibility)
 - ▶ data preprocessing (reproducibility)
 - ▶ task evaluation (comparability)

Usage Example

```
python main.py \  
  dataset=dwug_de_210 \  
  dataset/split=dev \  
  dataset/preprocessing=raw \  
  task/lscd_graded@task.model=apd_compare_all \  
  task/wic@task.model.wic=contextual_embedder \  
  task/wic/metric@task.model.wic.similarity_metric=cosine \  
  task.model.wic.ckpt=bert-base-german-cased \  
  task=lscd_graded \  
  evaluation=change_graded
```

Proof of Concept

- ▶ test two model alternatives on a common dataset (DWUG DE) under comparable conditions

Full graph representation

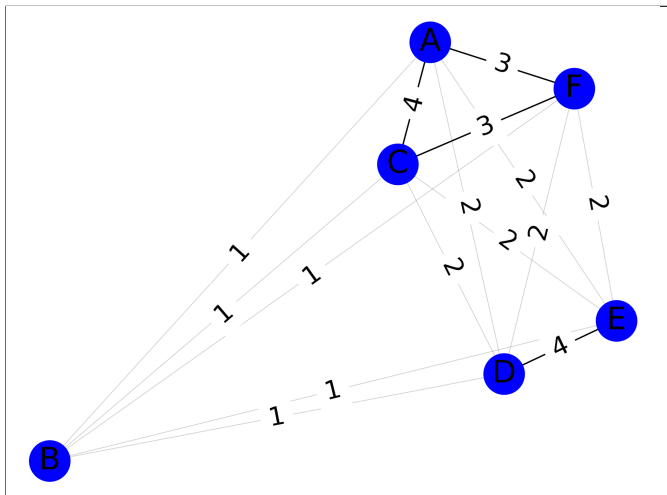
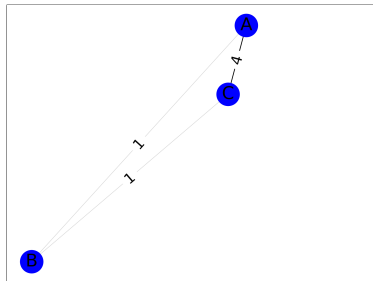
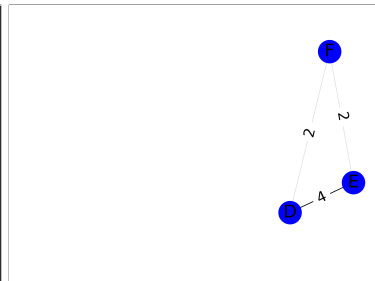


Figure 8: Word Usage Graph of English *arm*.

Time-wise subgraphs (EARLIER and LATER)



t_1 (EARLIER)



t_2 (LATER)

COMPARE subgraph

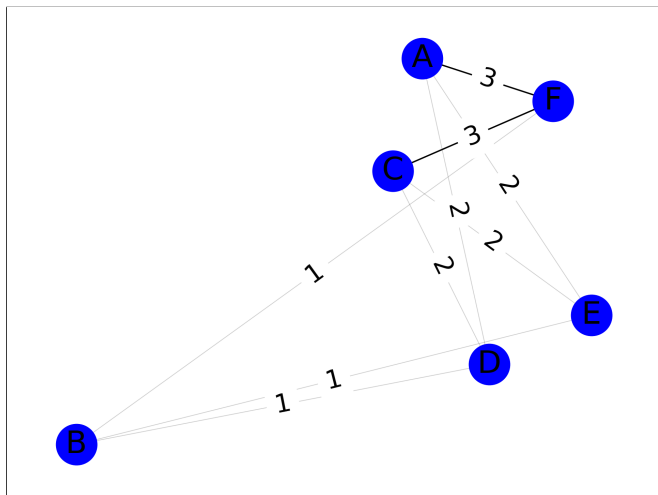


Figure 10: COMPARE subgraph of English *arm*.

Common graded change models

- ▶ **Average Pairwise Distance (APD)**: estimates the edge weights from COMPARE graph and takes their mean:

$$G(w) = \text{mean}(\text{COMPARE}) \quad (\text{e.g. Kutuzov \& Giulianelli, 2020})$$

- ▶ **DiaSense**: normalizes APD by weights from full graph:

$$G(w) = \text{mean}(\text{COMPARE}) - \text{mean}(\text{FULL}) \quad (\text{Beck, 2020})$$

Benchmark command APD

```
python main.py \  
  dataset=dwug_de_210 \  
  dataset/split=dev \  
  dataset/preprocessing=toklem \  
  task/lscd_graded@task.model=apd_compare_sampled \  
  task/wic@task.model.wic=contextual_embedder \  
  task/wic/metric@task.model.wic.similarity_metric=cosine \  
  task.model.wic.ckpt=bert-base-german-cased \  
  task=lscd_graded \  
  evaluation=change_graded
```


Benchmark command DiaSense

```
python main.py \  
  dataset=dwug_de_210 \  
  dataset/split=dev \  
  dataset/preprocessing=toklem \  
  task/lscd_graded@task.model=diasense_sampled \  
  task/wic@task.model.wic=contextual_embedder \  
  task/wic/metric@task.model.wic.similarity_metric=cosine \  
  task.model.wic.ckpt=bert-base-german-cased \  
  task=lscd_graded \  
  evaluation=change_graded
```

Result

Model	Run 1	Run 2	Run 3
APD	.63	.61	.63
DiaSense	.64	.55	.61

Table 3: Performance of model alternatives under comparable conditions on DWUG DE.

References I

- Beck, C. (2020). DiaSense at SemEval-2020 Task 1: Modeling sense change via pre-trained BERT embeddings. In *Proceedings of the 14th international workshop on semantic evaluation*. Barcelona, Spain: Association for Computational Linguistics.
- Kutuzov, A., & Giulianelli, M. (2020). UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th international workshop on semantic evaluation*. Barcelona, Spain: Association for Computational Linguistics.
- Schlechtweg, D. (2023). *Human and computational measurement of lexical semantic change* (Doctoral dissertation, University of Stuttgart, Stuttgart, Germany). Retrieved from <http://dx.doi.org/10.18419/opus-12833>