University of Stuttgart
Germany

DGfS Computerlinguistik-Tutorium

Lexical Semantic Change Detection

March 4, 2025

Dominik Schlechtweg

University of Stuttgart

Outline

Introduction

- Lexical Semantic Change Detection (LSCD)
  - goal: automate the analysis of changes in word meanings over time
- applications in historical linguistics, lexicography, but also social sciences ...
- obstacles for researchers: limited programming knowledge, missing standardization

→ this tutorial introduces two softwares aimed at simplifying the application of LSCD models:

1. DURel tool[1]
2. LSCD benchmark[2]

- particular view of semantic change (Schlechtweg, 2023)
- if you want to run the benchmark here in class, make sure to start the installation and one of the breaks
- these slides are available on my website[3]

[1] `https://durel.ims.uni-stuttgart.de`
[2] `https://github.com/Garrafao/LSCDBenchmark`
[3] `https://garrafao.github.io`

Schedule

Time	Topic
13:00 – 13:15	Introduction
13:15 – 14:00	Background
14:15 – 15:00	DURel tool
15:15 – 15:45	WUG datasets
16:00 – 16:45	LSCD benchmark
16:45 – 17:00	Closing

Background

- 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**!'

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**, 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.

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

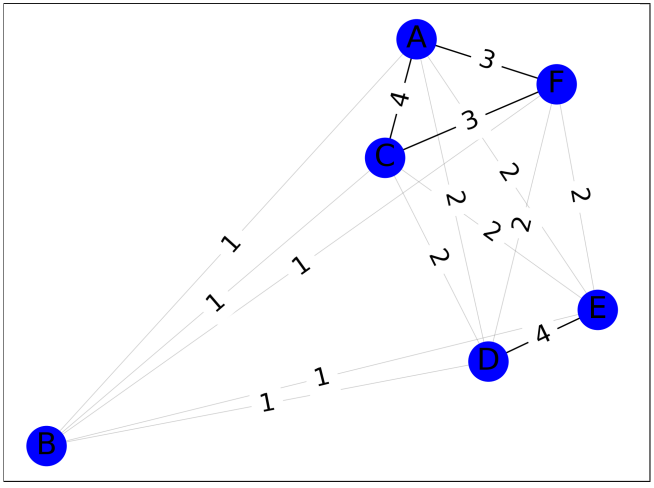


Figure 1: Word Usage Graph of English *arm*.

Clustering

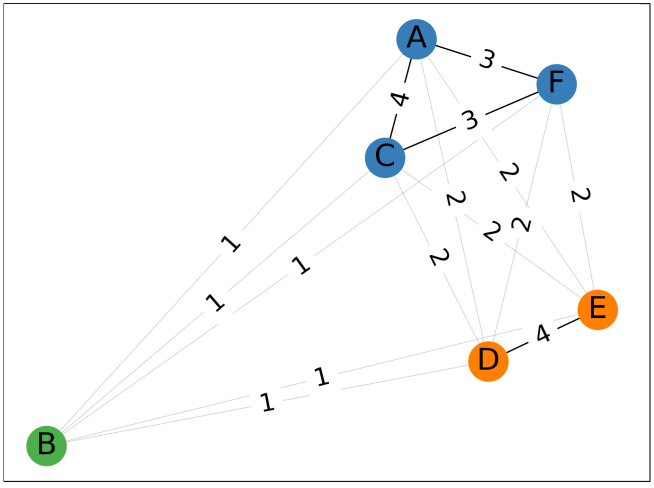
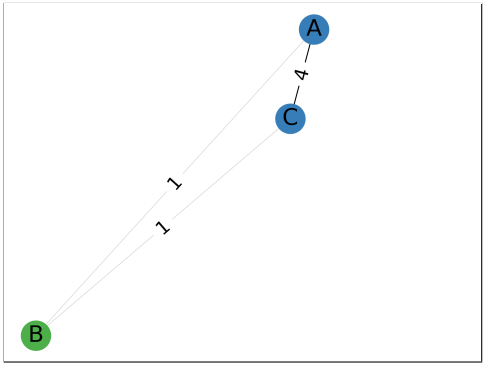
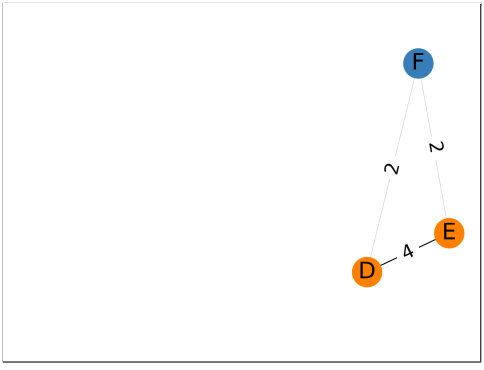


Figure 2: Word Usage Graph of English *arm*. $D = (3, 2, 1)$.

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*[4]

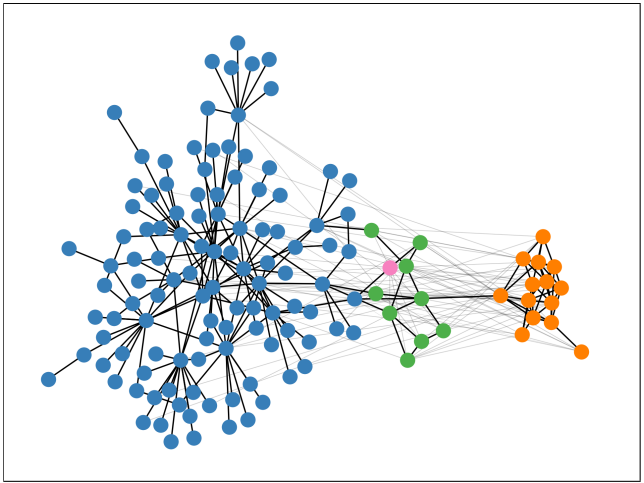


Figure 4: WUG of Swedish *ledning*.

[4]Datasets available at `https://www.ims.uni-stuttgart.de/data/wugs`

Example: Swedish *ledning*

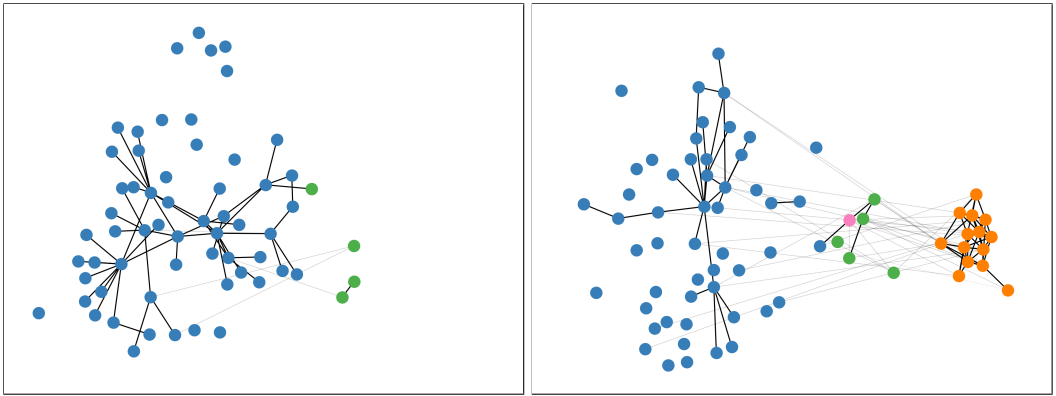


Figure 5: WUGs of Swedish *ledning*: subgraphs for 1st time period G_1 (left) and 2nd time period G_2 (right). $D_1 = (58, 0, 4, 0)$, $D_2 = (52, 14, 5, 1)$, $B(w) = 1$ and $G(w) = 0.34$.

Example: German *Eintagsfliege*

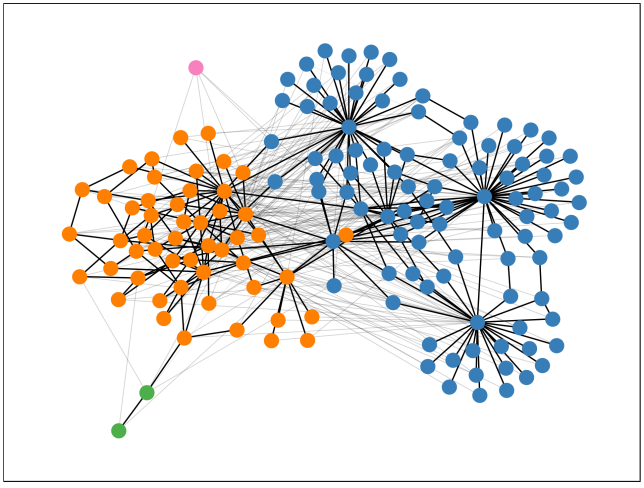


Figure 6: WUG of German *Eintagsfliege*.

Example: German *Eintagsfliege*

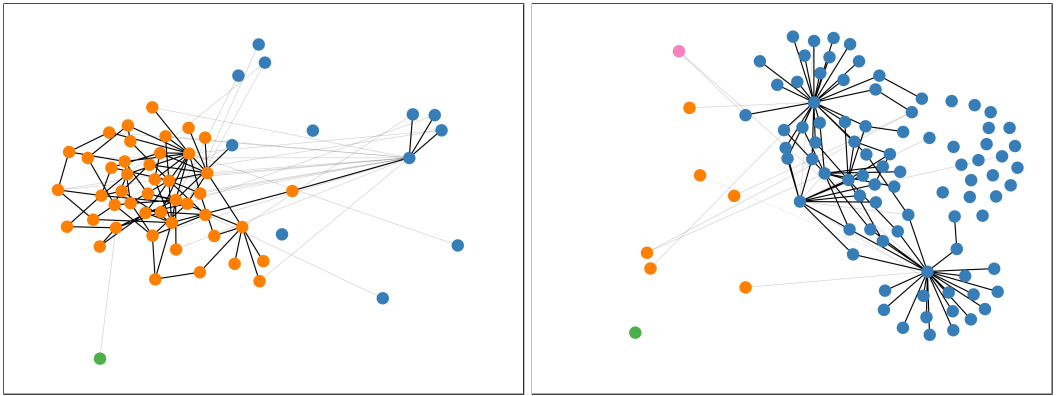


Figure 7: WUG of German *Eintagsfliege*: subgraphs for 1st time period G_1 (left) and 2nd time period G_2 (right). $D_1 = (12, 45, 0, 1)$, $D_2 = (85, 6, 1, 1)$, $B(w) = 0$ and $G(w) = 0.66$.

Summary of Annotation Steps

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

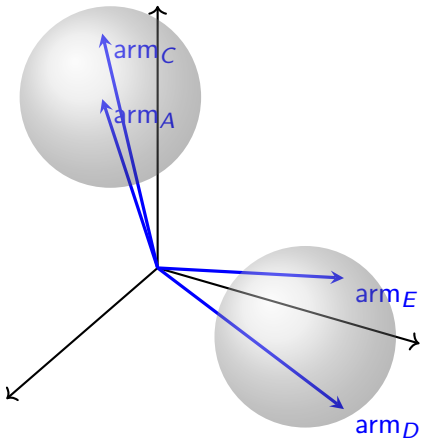
Summary of Annotation Steps with Tasks

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

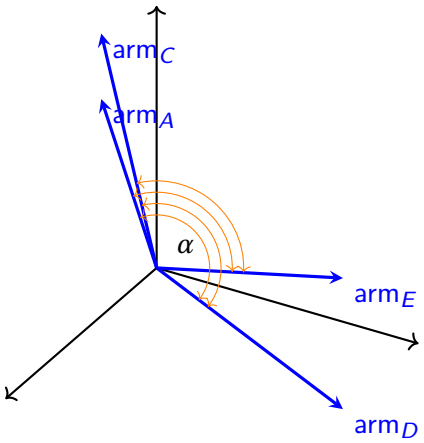
Computational Measurement of Lexical Semantic Change

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

Simple token-based Model



SOTA Model for graded change: APD



SOTA Model Components

- SOTA Models used for the different levels
  1. **semantic proximity**: DeepMistake, XL-Lexeme, GlossReader
     (Arefyev et al., 2021; Arefyev & Rachinskiy, 2021; Cassotti et al., 2023)
  2. **clustering**: Agglomerative, spectral, correlation, stochastic blockmodel
     (cf. Schlechtweg et al., 2024)
  3. **change measure**: Jensen Shannon Distance, Average Pairwise Distance
     (Kutuzov & Giulianelli, 2020; Lin, 1991)

DURel tool[5]

- demo
- register, upload, data inspection, annotation, visualization, computational annotation

[5] https://durel.ims.uni-stuttgart.de

WUG datasets[6]

- demo
- website, code repository, folder structure
- some examples

[6] https://www.ims.uni-stuttgart.de/data/wugs

graft

- **lemma**: graft
- **POS**: noun
- **dataset**: dwug_en_resampled
- **senses**[7]:
  - "shoot inserted into another plant," late 15c
  - "corruption," 1865, perhaps 1859

[7] etymonline.com

stab

- **lemma**: stab
- **POS**: noun
- **dataset**: dwug_en_resampled
- **senses**[8]:
  - mid-15c., stabbe, "wound produced by stabbing," from or related to stab (v.).
  - The meaning "act of stabbing, thrust or blow with the point of a weapon" is from 1520s.
  - The meaning "a try" is recorded by 1895, American English, probably on the notion of "a vigorous thrust."
  - The verb was long used in reference to a wound given treacherously; stab in the back in the figurative sense "treacherous deed" is attested by 1881; the verbal phrase in the figurative sense is from 1888.

[8] etymonline.com

Ohrwurm

- **lemma**: Ohrwurm
- **POS**: noun
- **dataset**: dwug_de
- **senses**[9]:
  - kleines bis mittelgroßes, längliches, meist geflügeltes, nachtaktives Insekt mit längeren Fühlern und mit einem Paar Zangen am Ende des Hinterleibes, das sich vor allem krabbelnd fortbewegt, von Pflanzen ernährt und auch andere Insekten o.Ä. jagt
  - [übertragen] Musikstück mit einer eingängigen, einprägsamen Melodie, das einem Hörer für einen längeren Zeitraum in Erinnerung bleibt

[9] dwds.de

abgebrüht

- **lemma**: abgebrüht
- **POS**: adjective
- **dataset**: dwug_de
- **senses**[10]:
  - unempfindlich

[10] `dwds.de`

Hands-on

- **Task**: Find out whether and how word **ausspannen/record** has changed its meaning over time. Build two groups: one group inspects the raw data, one the human annotation.

ausspannen

- **lemma**: ausspannen
- **POS**: verb
- **dataset**: dwug_de
- **senses**[11]:
  - a) etw. zwischen zwei Punkten straff spannen, b) etw. spannend ausbreiten, c) [gehoben, übertragen] ⟨sich ausspannen⟩ sich ausdehnen
  - etw. Eingespanntes lösen
  - [umgangssprachlich] ... a) jmdm. etw. mit (schließlich erlangter) Einwilligung des Besitzenden wegnehmen, b) jmdm. jmdn. abspenstig machen
  - längere Zeit von der Arbeit ausruhen, sich erholen

[11] dwds.de

record

- **lemma**: record
- **POS**: noun
- **dataset**: dwug_en_resampled
- **senses**[12]:
  - c. 1300, "testimony committed to writing, fact or condition of having been recorded,"
  - "a written account of some fact, event, or proceeding for the purpose of preserving the memory of it" is from late 14c.
  - "disk on which sounds or images have been recorded" is attested from 1878
  - "best or highest official achievement in a sport, activity, etc." is by 1883
  - The journalist's phrase on the record is attested from 1900; adverbial phrase off the record "confidentially" is attested from 1906.

[12] etymonline.com

The LSCD Benchmark

- Lexical Semantic Change Detection (Schlechtweg, 2023)
- **heterogeneity** and **modularity** in models, datasets and tasks

→ create one repository[13] standardizing model component combinations, dataset preprocessing and evaluation

[13] `https://github.com/Garrafao/LSCDBenchmark`

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
```

Full graph representation

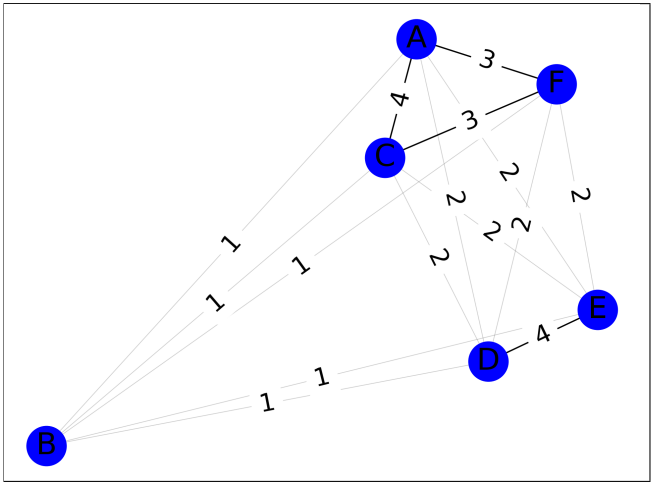
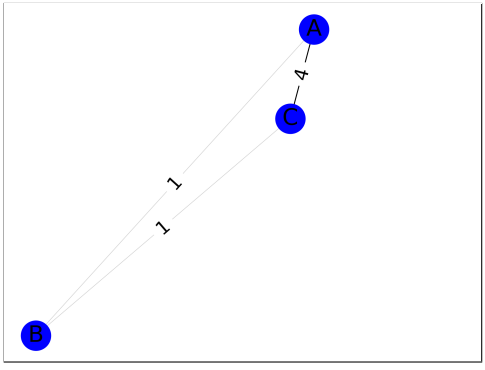
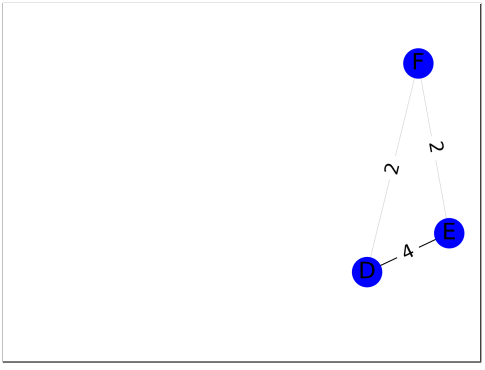


Figure 8: Word Usage Graph of English *arm*.

Time-wise subgraphs (EARLIER and LATER)



t_1 (EARLIER)



t_2 (LATER)

COMPARE subgraph

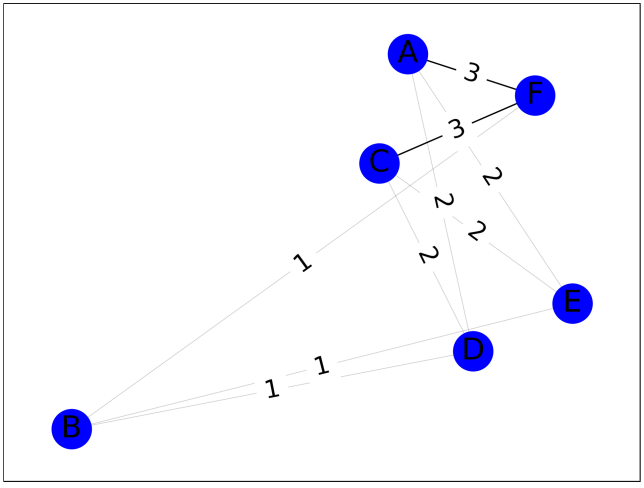


Figure 10: COMPARE subgraph of English *arm*.

Common graded change models

- **Average Pairwise Distance** (APD): estimates the edge weights from COMPARE graph and takes their mean: $G(w) = mean(COMPARE)$ (Kutuzov & Giulianelli, 2020)
- **DiaSense**: normalizes APD by weights from full graph: $G(w) = mean(COMPARE) - mean(FULL)$ (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
```

Hands-on

- demo
- implementation details, installation, modeling levels, running model on dataset, adding new models
- **Task**: Implement APD

Upcoming

- How do models perform on new data sets?
- Are current results for SOTA models reproducible on cleaned data?
- Can we find better measures for graded change than APD?
- Can clustering on optimized WiC models improve results on binary change?
- How do current models for binary change perform in scenarios where correlation is low with graded change?
- How do current models for graded change perform in high-polysemy scenarios?
- How do model hyper-parameters generalize between data sets?

Closing

- evaluation: `https://shorturl.at/AHA9H`
- Thank you for participating!

References I

Arefyev, N., Fedoseev, M., Protasov, V., Homskiy, D., Davletov, A., & Panchenko, A. (2021). DeepMistake: Which senses are hard to distinguish for a word-in-context model. In (Vol. 2021-June, pp. 16–30). Retrieved from https://www.dialog-21.ru/media/5491/arefyevnplusetal133.pdf

Arefyev, N., & Rachinskiy, M. (2021). Zero-shot cross-lingual transfer of a gloss language model for semantic change detection. In (Vol. 2021-June, pp. 578–586). doi: 10.28995/2075-7182-2021-20-578-586

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.

Cassotti, P., Siciliani, L., de Gemmis, M., Semeraro, G., & Basile, P. (2023, July). Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61th annual meeting of the association for computational linguistics*. Online: 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.

Lin, J. (1991, 01). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, *37*, 145–151.

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

Schlechtweg, D., Zamora-Reina, F. D., Bravo-Marquez, F., & Arefyev, N. (2024). Sense through time: Diachronic word sense annotations for word sense induction and lexical semantic change detection. *Language Resources and Evaluation*. Retrieved from https://doi.org/10.1007/s10579-024-09771-7