NorDiaChange: Diachronic Semantic Change Dataset for Norwegian

Andrey Kutuzov[1], Samia Touileb[2], Petter Mæhlum[1], Tita Ranveig Enstad[1], Alexandra Wittemann[1]
[1] University of Oslo, [2] University of Bergen

NorDiaChange

- Manually annotated following the DWUG methodology (Schlechtweg et al., 2021).
- Fully compatible with datasets for other languages.
- `https://github.com/ltgoslo/nor_dia_change`

Corpora used

- **NBDigital** from the National Library of Norway: historical corpus of over 26,000 books, reports, and news articles.
- **Norwegian newspaper corpus** (Norsk Aviskorpus or NAK).

We provide two independent datasets (can be used as train/test interchangeably):

Subset 1: 1929-1965 VS 1970-2013

- Important historical periods for Norway: pre- and post-war, pre- and post-oil.
- Samples from NBDigital corpus.

Subset2: 1980-1990 VS 2012-2019

- Changes caused mostly by technological advances.
- Samples for the 1st period from NBdigital, for the 2nd from NAK.

Period	Words	Documents
1929 – 1965	57 mln	959
1970 – 2013	175 mln	4,209
1980 – 1990	43 mln	1,115
2012 – 2019	649 mln	1,763,843

Total number of words and documents in both time periods of Subsets 1 and 2.

Target words and annotations

- **Nouns** manually selected, 40 in total, based on linguistic intuition.
- added a random filler word for each selected target word.
- DURel framework (Schlechtweg et al., 2018) with accompanying web service.
- Annotators judge *word usage pairs* on the graded scale of *sense relatedness*:
  - *identical*, *closely related*, *distantly related*, *unrelated*, cannot decide
- Resulting **word usage graphs** are clustered to infer **diachronic changes in sense distributions**.
- **Graded change score** is Jensen-Shannon distance (JSD) between these distributions.
- **Binary change (sense lost/sense gained)** is inferred with simple heuristics.

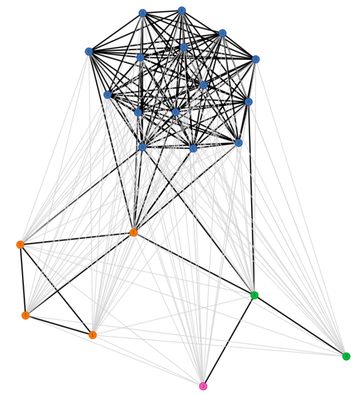
Dataset	Words	$\|U\|$	JUD	SPR	KRI
Subset 1	40	21	12,727	0.77	0.76
Subset 2	40	21	14,003	0.71	0.67

$|U|$: avg. num. usages sampled for a word. JUD: total num. of judgements. SPR: weighted mean of pairwise Spearman ρ correlations between annotators. KRI: Krippendorff's α inter-rater agreement.

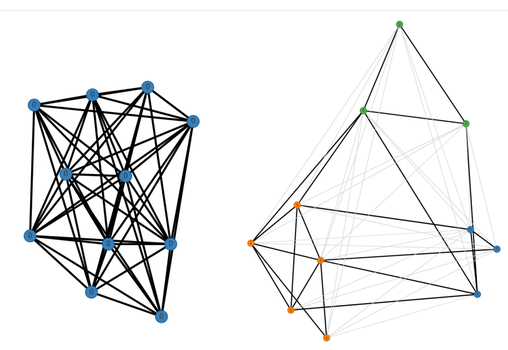
Word	Graded	Sense gain	Sense lost
Subset 1 (1929-1965 VS 1970-2013)			
plattform	0.87	1	1
leilighet	0.80	0	1
horisont	0.64	1	1
mål	0.60	0	1
bølge	0.60	1	0
Subset 2 (1980-1990 VS 2012-2019)			
stryk	1.00	1	1
kanal	0.73	0	1
kode	0.73	1	1
oppvarming	0.72	1	0
innstiling	0.66	1	0

Top changed words in NorDiaChange.

Word usage graphs



WUG for *innstilling*, 2 time periods. Senses: *'ruling'* (0), *'attitude'* (1), *'setting'* (2).



Time-specific WUGs for *oppvarming*. Left: the only sense of *heating* (0). Right: senses of *heating* (0), *global warming* (1), and *warm-up* (2).

References

(Schlechtweg et al., 2018) Schlechtweg, D., Schulte im Walde, S., and 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, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

(Schlechtweg et al., 2021) Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., and McGillivray, B. (2021). DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.