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

Norway

andreku@ifi.uio.no, Samia.Touileb@uib.no, pettemae@ifi.uio.no, titare@ifi.uio.no, alexankw@ifi.uio.no

Abstract

We describe *NorDiaChange*: the first diachronic semantic change dataset for Norwegian. *NorDiaChange* comprises two novel subsets, covering about 80 Norwegian nouns manually annotated with graded semantic change over time. Both datasets follow the same annotation procedure and can be used interchangeably as train and test splits for each other. *NorDiaChange* covers the time periods related to pre- and post-war events, oil and gas discovery in Norway, and technological developments. The annotation was done using the DURel framework and two large historical Norwegian corpora. *NorDiaChange* is published in full under a permissive licence, complete with raw annotation data and inferred diachronic word usage graphs (DWUGs).

Keywords: semantic change detection, semantics, datasets, Norwegian

1. Introduction

In this paper, we present *NorDiaChange*: the first dataset of diachronic semantic change on the lexical level for Norwegian. Such datasets are required to evaluate lexical semantic change detection systems or contextualized embeddings in general, but can also be useful for historical linguists. *NorDiaChange* is naturally accompanied with publicly available historical corpora. These historical corpora were used to produce *NorDiaChange* via a meticulous manual annotation effort following the DWUG (Diachronic Word Usage Graphs) methodology (Dominik Schlechtweg and Sabine Schulte im Walde, 2021): in short, it means that the annotators are yielding judgements about how semantically similar a word x is in sentence pairs shown to them. As such, *NorDiaChange* is fully compatible with datasets for other languages used, for example, in the SemEval'2020 shared task on semantic change detection (Schlechtweg et al., 2020). However, it is different from most of them in that *NorDiaChange* features two independent Norwegian datasets dubbed here 'Subset 1' and 'Subset 2'. Subset 1 deals with semantic change occurring between the period of 1929-1965 and the period of 1970-2013. Subset 2 focuses on the time periods of 1980-1990 and 2012-2019. Since the annotation procedure was exactly the same for both subsets, they can be used as train-test splits for each other.

NorDiaChange is published in full[1] with all the raw annotation judgements so that any preferred scoring workflow can be applied to it. However, we stick to the standard DWUG scoring procedure described below and provide a graded and a binary change score to each target word.

The rest of the paper is structured as follows. In section 2, we put our annotation effort in the context of prior work on semantic change detection. Section 3 describes the historical corpora of Norwegian we used to sample sentences for annotation. In section 4, we explain how we selected the target words to annotate. Further on, section 5 describes the annotation process itself. In section 6, we conduct qualitative linguistic analyses of the annotation results and sanity-check the output of the scoring algorithms we used. Finally, in section 7, we conclude and discuss possible future avenues for our work.

2. Related work

Studying semantic change is a venerable field in linguistics; see (Bloomfield, 1933) and (Blank and Koch, 1999), among many others. However, in natural language processing, the topic of automatically tracing semantic change received comparatively little attention until the advent of easy-to-use distributional representations of lexical meaning (word embeddings). Kulkarni et al. (2015), Hamilton et al. (2016a), Hamilton et al. (2016b) and others have shown the potential that distributional semantics has for this task. We refer interested readers to comprehensive reviews on the topic: one can mention (Kutuzov et al., 2018) and (Tahmasebi et al., 2021), among others.

In order to compare different semantic change detection methods, one needs access to high-quality evaluation datasets. Although attempts to create such resources started as early as at least in 2011 (Gulordava and Baroni, 2011), they were diverse, not standardised, and suffered from sparse and biased data selection.

Currently, the mainstream approach to avoid this problems is to employ *graded* contextual word meaning annotation, with the DURel framework (Schlechtweg et al., 2018) being the most prominent example. In it, annotators are shown contextualized word usage pairs and asked to judge semantic similarity of two usages for the same word on a graded scale. After that, a change score is inferred from these judgements (see Section 5 for more details). SemEval'2020 shared task on unsu-

---

[1] https://github.com/ltgoslo/nor_dia_change

pervised semantic change detection employed datasets annotated within this framework, further developing it to include smart sampling of usage pairs and clustering word usages based on their relations to each other in a usage graph (Schlechtweg et al., 2020).

A large trove of diachronic *word usage graphs* annotated this way exists for several languages (English, German, Swedish, Latin) (Schlechtweg et al., 2021), but this set lacks Norwegian. *NorDiaChange* is aimed at filling in this gap, by being fully compatible to the existing datasets (produced using the same procedure). Additionally, we are following (Rodina and Kutuzov, 2020) and (Kutuzov and Pivovarova, 2021b) in providing *several* independent datasets for a particular language (in our case, Norwegian). These *subsets* are different in their target time periods and word lists, but are annotated in the same way and by the same team. This potentially allows to develop *supervised* semantic change detection systems which can be trained or fine-tuned on one subset and then evaluated on another. Systems using fine-tuning instead of 'zero-shot' approaches turned out to be the best in the recent semantic change detection shared task for Russian (Kutuzov and Pivovarova, 2021a), and so we consider this possibility to be crucial for our *NorDiaChange* dataset as well.

3. Corpora used

The underlying corpora of our work are the NBdigital[2] corpus from the National Library of Norway, and the Norwegian newspaper corpus (Norsk Aviskorpus or NAK[3]). Both corpora are freely available, and can be downloaded from the website of the National Library of Norway. The NBdigital is a historical corpus containing a collection of over 26,000 books, reports, and news articles from the public domain. The texts cover various genres, written in various languages including both Norwegian forms (Bokmål and Nynorsk). In addition, each text has a list of metadata information, as e.g. author, OCR confidence, and date of publication. However, due to the use of OCR, not all of the texts are of acceptable quality. We therefore had to first filter out all texts with an OCR confidence below 70%, and thereafter cleaned the non Norwegian Bokmål texts from the Bokmål collection of NBdigital.

The content of NBdigital is dated up to 2013. In order to get more recent data that can reflect more modern language use, we have also selected news articles from the NAK corpus. We therefore only use articles published between 2012 and 2019 from NAK.

We decided to divide the corpus into two subsets. The first subset compares the time periods of 1929-1965 and 1970-2013. In the second subset, we look at the time period of 1980-1990 compared to 2012-2013. In

Period	Words	Documents
1929 – 1965	57 mln	959
1970 – 2013	175 mln	4,209

Table 1: Total number of word tokens and documents in Subset 1.

Period	Words	Documents
1980 – 1990	43 mln	1,115
2012 – 2019	649 mln	1,763,843

Table 2: Total number of word tokens and documents in Subset 2.

what follows, we give the arguments behind this decision, and describe the content of each subset.

3.1. Subset 1: 1929-1965 VS 1970-2013

This subset captures two important historical time periods for Norway. The pre- and post-war periods have affected Norwegians' standard of living, and therefore their language use. Higher living standards and better economy after the 1960s made more people, traditionally farmers, move to bigger cities[4]. By the advent of the 1970s a consumer culture was established and new technology entered Norwegian homes. All of these changes impacted the language use by both introducing new words to the vocabulary, and adding new senses to pre-existing words. The time period between the 1970s and 2013 also introduced quite many technologically related words and senses, and as society developed, the language arguably followed the trends[4].

In this subset, both time periods are extracted from the historical NBdigital corpus. Table 1 shows the total number of words and documents in both time periods of Subset 1.

3.2. Subset2: 1980-1990 VS 2012-2019

This subset contains shorter time periods than in Subset 1, but we still expect shifts in word usages. The changes between the two periods of 1980-1900 and 2012-2019 can be caused both by linguistic and cultural factors. Most of the changes are expected to be within technology, and the changes in language use are certainly mostly in vocabulary additions. However, many words related to technological advances were added as new senses to pre-existing Norwegian words.

The first time period is extracted from NBdigital, while the second one from 2012 to 2019 comprises texts from the NAK corpus. The language use in NAK might be different from NBdigital, as we expect news texts to contain more 'modern' senses of words that have shifted from the previous time period. In Table 2, we

give an overview of the total number of words and documents in each time period of Subset 2.

4. Choosing target words

We manually selected target words that we believe can have undergone semantic changes during the periods of Subset 1 and Subset 2. This was based on linguistic intuition of the authors as native Norwegian speakers and on existing linguistic work, similar to what has been done by Rodina and Kutuzov (2020) and Schlechtweg et al. (2021). Another option would be to look at diachronic dictionaries, but Vikør (2006) notes that although some dictionaries make note of antiquated senses for some lemmas, there are no truly diachronic dictionaries for Norwegian, with the exception of etymological dictionaries.

For each of the selected target words, a filler word was added to the word list. Filler words were randomly sampled from the corresponding corpora; their frequency percentiles in both earlier and later time periods had to be similar to those of the corresponding target words (in this, we followed the current best practices). The purpose was to ensure that corpus frequency dynamics alone cannot be used to solve the dataset (frequency changes often accompany semantic changes).

Indeed, there are no statistically significant correlations between frequency differences (IPM-normalised) and annotated change scores. For Subset 1, Spearman ρ is -0.06 ($p = 0.70$), for Subset 2 it is 0.11 ($p = 0.49$). It means that *NorDiaChange* does not just list words which sharply changed their corpus frequency together with semantic shifts. It is balanced with regards to word frequencies, and the systems aiming to approximate the scores in *NorDiaChange* based on corpus data must take into account more than that.

Note that the filler words were manually checked to make sure that they are valid Norwegian lemmas not immediately associated with any known diachronic semantic shift. However, after the annotation, we found out that some of the filler words actually did experience some historical change; see section 6 below. All our target and filler words are nouns and we discarded words belonging to other parts of speech.

The initial target word lists for both subsets (without the fillers) contained 20 words each (40 target words in total). These words are expected to have semantically shifted between 1929-1965 and 1970-2013 for Subset 1, and between 1980-1990 and 2012-2019 for Subset 2. For some of the words, we also did a dictionary check to see if the word seemed to have lost senses in different time periods. The full list (including senses) can be found in the project GitHub repository.

5. Annotation process

The data was annotated using the DURel framework (Schlechtweg et al., 2018) and the accompanying web service.[5] The annotators were presented with *word usage pairs*: two snippets of text, both containing the same target word. The annotation task is to decide the semantic similarity of the target word in the two different contexts, using the following graded scale:

- senses are *identical*: the word usage pair is annotated with 4

- senses are *closely related*: 3

- senses are *distantly related*: 2

- senses are *unrelated*: 1

- the annotator is unsure and cannot decide the semantic similarity: 0

Thus, annotators are judging usage pairs on a semantic proximity scale, avoiding any manual definitions of word senses. An example of semantically unrelated usages is the word *ris* in the following usage pair:

1. *Sidene pagineres i hver bok fra 1—500: Papiret koster pr. ris à 500 ark kr. 32,80, hvorpå beregnes 20 pet. fortjeneste.*
   ('The pages are numbered from 1-500 in each book: The paper costs for each **ream** of 500 sheets NOK 32,80, to which a 20% profit is calculated.')

2. *I saltvann produseres primært grønn tang til en samlet verdi av $ 440 mill. (1973). Denne tangen spises til ris som en slags grønnsak.*
   ('In seawater green seaweed is primarily produced for a total value of $ 440 million (1973). This seaweed is eaten with **rice** as a type of vegetable.')

In the first sentence, *ris* refers to a unit of quantity of paper, while in the second sentence, it refers to rice. The annotator is expected to yield the 1 judgement.

Contextualised word usages are essentially sentences containing at least one target word (can be a filler). These sentences were sampled from the historical corpora described in section 3 in the following way. First, we lemmatized and POS-tagged our corpora using UD-Pipe (Straka and Straková, 2017). This was required to be able to find inflected forms of the target words. From the processed corpora, we extracted all sentences containing at least one token with a target word as its lemma (we filtered out target words not tagged as a noun). The extracted sentences (both raw text and processed versions with lemmas and POS tags) were stored as time period marked JSON files for further usage: one file per target word / time period.

From each of these files, we randomly sampled 11 sentences as representatives of a particular target word usage in a particular time period. Before sampling, the sentences were de-duplicated, and we discarded sentences containing the '^' character (in most cases, it

---

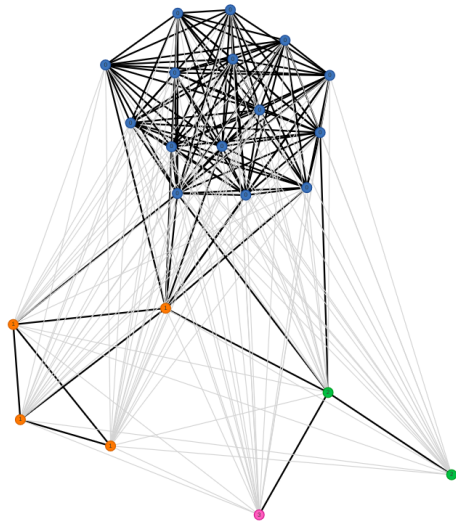[5] https://www.ims.uni-stuttgart.de/data/durel-tool

Figure 1: Word usage graph for *innstilling*, time periods 1980-1990 and 2012-2019.

signalled heavy OCR errors). This means that for every target word, we had 22 total usage examples (11 from the earlier and 11 from the later time period). In the rare cases when there were less than 11 occurrences of the target word[6], all the existing occurrences were used.

11 might seem to be a small sample: Schlechtweg et al. (2021) sampled 100 usages for each target word / historical corpus. We are aware of the limitations of our sample size, and that we cannot guarantee that all senses are covered. However, one should keep in mind that not all possible usage pairs are getting annotated anyway (see more on that below). In the end, *NorDiaChange* contains 26, 730 annotator judgements, which is comparable to the existing datasets for English (29, 000) and Swedish (20, 000).

The sampled sentences were fed to the DURel web service which was responsible for randomly selecting usage pairs to present to the annotators. More details on that can be found in (Schlechtweg et al., 2021). In short, the aim here was to spend as little annotation effort as possible to construct a well-connected diachronic word usage graph (DWUG), where nodes are usages (sentences) and edges between them are weighted with annotator's judgements. Thus, sentences where a target word is used in roughly the same sense (4 judgements), are naturally grouped into a 'sense cluster'. See, for example, DWUG for the Norwegian word *instilling* in figure 1. Node colours correspond to automatically inferred sense clusters, edge thickness corresponds to the median annotators' judgements. The blue cluster contains sentences where *instilling* is used in the older sense of 'recommendation', and they come from both 1980-1990 (earlier) and the 2012-2019 (later) time periods. However, the orange, green, and lilac clusters contain exclusively sentences

---

from the later time period, with *instilling* used in the senses of 'setting' (as in 'account settings') or 'attitude'. Thus, this DWUG constitutes a case of a word gaining new senses diachronically.

NorDiaChange is published with full raw annotation data, so it is possible to infer sense grouping from human similarity judgements on usage pairs in any preferred way. However, we followed the standard workflow from (Schlechtweg et al., 2021) and employed their code to produce clustered DWUGs. After the sense clusters are inferred, it becomes possible to compare sense distributions across time (normalised to become probability distributions). The graded change score was the main score we assigned each word in *NorDiaChange*. It is calculated as Jensen-Shannon distance (JSD) between the probability distributions of senses in the earlier and the later time periods (Schlechtweg et al., 2019; Giulianelli et al., 2020).

To continue the example with *instilling* on figure 1, there are four sense clusters in total. In the earlier time period, all 11 usages belong to the first cluster, so the usage distribution is $[11, 0, 0, 0]$ or $[1, 0, 0, 0]$ when converted to a probability distribution. In the later time period, the usage distribution is $[4, 4, 2, 1]$ (only 4 of the usages belong to the first cluster), and after converting to a probability distribution, it is $[0.364, 0.364, 0.182, 0.091]$. The Jensen-Shannon distance between $[1, 0, 0, 0]$ and $[0.364, 0.364, 0.182, 0.091]$ equals to 0.655 which is the degree of diachronic change that the word *instilling* experienced between 1980-1990 and 2012-2019.

Note that due to essentially random sampling of sentence pairs, some potential groupings of nodes (sentences) might end up unconnected with edges. This can in theory negatively influence the clustering. This is why we also calculated the number of uncompared cluster combinations for each target word. It is zero in almost all the cases, with the exception of one word in the Subset 1 (*bølge*) and two words in the Subset 2 (*kanal* and *sky*).

In addition, we provide binary change scores, where each word is assigned a 1 label if it gained or lost any senses between two time periods or a 0 label if its senses remained stable. In determining these scores, we again followed the approach from (Schlechtweg et al., 2020). A word is considered to experience a binary change, if at least one sense was attested at most k times in one time period and at least n times in another time period. n and k here are manually defined hyperparameters which are needed to filter out insignificant fluctuations in sense frequencies. We kept the default values of $k = 1$ and $n = 3$. This means that, for example, having an entirely novel sense cluster with two usages in it is not enough to assign a 1 binary change score (since $2 < 3$). However, in the case of *instilling*, *NorDiaChange* marks it as binary changed, since its second sense is represented with 4 usages in the later time period ($4 \geq 3$) and 0 usages in the earlier time

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
Swedish	40	168	20,000	0.57	0.56
Russian	99	60	8,879	0.56	0.55

Table 3: Descriptive statistics for the full *NorDiaChange* subsets. $|U|$ is the average number of usages sampled for a word from the corpora. JUD is the total number of annotators' judgements for a subset. SPR is the weighted mean of pairwise Spearman ρ correlations between different annotators, and KRI is the value of Krippendorff's α inter-rater agreement.

period $(0 \leq 1)$.

5.1. Annotators

NorDiaChange was annotated by three native speakers of Norwegian, all of whom hold bachelor degrees in either linguistics or language technology. They received identical guidelines which can be found at `https://github.com/ltgoslo/nor_dia_change/blob/main/guidelines.md`. Several reconciliation meetings were held between the annotators and project managers to make sure that all annotators understand the guidelines in the same way, and to clear up inconsistencies in the data. Each annotator produced from $9,167$ to $10,584$ judgements in total and was paid about $13,000$ Norwegian Krone (about $1,300$ euros) for their job after taxes.

5.2. Annotation results

Each usage pair received 1.8 annotator's judgements on average, so possible errors or misunderstandings of one annotator could be compensated by another. The most important descriptive statistics for both *NorDiaChange* subsets are given in Table 3; more is available at our repository. For comparison, we also provide the same statistics for the Swedish dataset from (Schlechtweg et al., 2021) and the Russian *RuShiftEval-2* dataset from (Kutuzov and Pivovarova, 2021b).

Table 4 shows top 5 most changed words for both *NorDiaChange* subsets. The 'graded' column gives the values of the graded change score (calculated with JSD). The 'sense gain' and 'sense lost' columns demonstrate a binary label depending on whether a word acquired or lost at least one sense based on the binary change score calculated with the same default values of $k = 1$ and $n = 3$.

It is interesting to note that one of those words in Subset 1 (*horisont*) and three in Subset 2 (*stryk*, *oppvarming* and *innstilling*) were filler words, randomly sampled from the corpora: that is, we did not intentionally came up with these words as 'changed'. This is normal (the same was observed, for example, by Kutuzov and Pivovarova (2021b)) and demonstrates that mining real textual data can sometimes yield unexpected but

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

Table 4: Top changed words in *NorDiaChange*.

still valid insights. However, in general human annotations correspond well to our original intuitions: if we assign the value of 1.0 to the words originally chosen as target ones and 0.0 to the words originally sampled as fillers, then both subsets show highly statistically significant ranked correlations of these values and the graded change scores. For Subset 1, Spearman ρ in this case is 0.38 ($p = 0.01$) and for Subset 2, it is 0.40 ($p = 0.02$).

As it was expected based on prior work, some target words received disproportionately many 0 annotations, for different reasons.[7] In some of those cases, a target word has a verbal homonym which was erroneously tagged as a noun by UDPipe. All occurrences where the target word was a verb in a word usage pair were annotated with 0. For the target word *vert* 'host' from Subset 2 , 27% of the examples were actually the verb *verte* 'become'. This resulted in the word usage graph for 1980-1990 only containing three nodes (usages with high proportion of zero judgements are removed from the graph automatically). The annotation of the word *tap* 'loss' from Subset 2 also yielded small word usage graphs due to many 0 annotations. This was in part because 9% of the context examples were actually the verb *tape* 'lose'.

On the other hand, the word *fil* 'file' from Subset 1 had many 0 annotations due to noisy data. There were so many examples where it was impossible to interpret the context, that the word usage graph for 1929-1965 only had three nodes, all of them in the same cluster.

We also had some words where the team didn't fully agree with the produced results. The usage graph for the word *mening* 'meaning, opinion' from Subset 1 clustered together senses that should not have been in the same cluster. As a result, the scoring algorithm also marked this word as binary changed, contrary to

---

[7]Overall, there were 548 zero judgements in Subset 1 and 286 in Subset 2 (Subset 1 suffers much heavier from OCR errors).

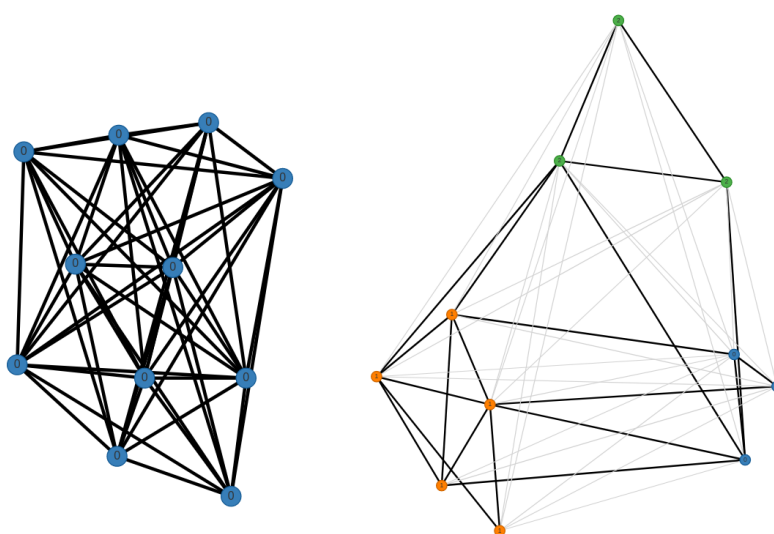**1980-1990**   **2012-2019**

Figure 2: The two word usage graphs for *oppvarming* aligned. The left graph shows the earlier time period (1980-1990), with the only sense of 'heating' (0). The right graph shows later time period (2012-2019), with the three clusters referring to 'heating' (0), 'global heating' (1) and 'warm-up' (2).

the opinion of the annotators. During a reconciliation meeting, the team agreed that the nuances in the word semantics were difficult to annotate due to the abstract nature of the word.

The annotation of the word *damp* 'steam' from Subset 1 yielded a somewhat 'strange' word usage graph for the time period 1929-1965. The graph contains four clusters, but two of these have the same sense (steamboat) and should have been grouped together into one cluster.

For the word *mus* 'mouse' from Subset 2, the annotation showed that the sense meaning computer mouse had disappeared, but all native speakers agree that this sense is still very much present in the language. See section 6 for more details. We decided to mark these problematic words as questionable, and publish both a 'clean' dataset with 37 target words in each subset, and the full set with all 40 annotated words in each subset, including the six 'questionable' ones.

6. Meaning change in the Norwegian language of the 20th century

In order to qualitatively analyse the binary and graded change scores produced within *NorDiaChange*, we consulted the annotators and authors whose native language was Norwegian. Their intuitions are the basis of these evaluations. The annotators consulted the graphs for each period, as in Figure 2, and looked at the sense clusters to examine which senses had been clustered in which ways. However, it can be difficult, even for native speakers, to personally judge the semantic graphs. In an attempt to mitigate native speaker bias, we believe that looking at frequencies over time can be one

way to get additional insight into the DWUGs. The National Library of Norway has released the National Library N-gram (NB N-gram) service, which allows users to check the corpus frequencies of an n-gram between 1810 and 2013. The data is similar to what was used for this annotation effort, with the exception of the NAK-data for the latter part of subset 2.[8] We use this service to highlight certain insights. The following discussion is based on the 'clean' dataset.

The annotators all agree on the binary results, but there are parts of the sense graphs that do not necessarily fit the annotators' intuition completely. One case is when a sense is overly specific. An example of this can be seen for the word *oppvarming* 'heating' (figure 2), where in the later period of 2012-2019, the new sense '(global) warming' contributes to about 5/8 of the graded change score. However, this sense is closely related to the 'heating' sense, and might not be seen as separate. It seems like the graph is sensitive to what might be perceived as different types of metaphorical usage. This is not necessarily an error, but it is worth noting that this can potentially expand the numbers of senses considerably. Another example of this is *bølge*, (see Figure 3) whose graph has as many as 6 senses. 5 of these are metaphorical: waves of gratitude, waves of battle, waves of hair, waves of grain, waves of song.

The other case is when two senses are conflated. This can happen both within the sense graph for one period, or across time periods for senses that are considered the same. One example is the graph for *skjerm*, in which it seems like at least one of the examples for the 'com-

---

[8]https://www.nb.no/sprakbanken/
ressurskatalog/oai-nb-no-sbr-70/

puter screen' sense, based on its sense in the later time period, also contains a sense of 'canvas used when taking a photo', or something related. This could also be due to difficulties during annotation, where annotators report that some senses are more difficult to judge from context than others. On the other hand, we used the default settings for graph clustering; tuning them may yield other groupings, more or less specific.

Another problem is frequency of senses in both corpora. Although the annotators agreed that it seems likely that the frequency of the sense 'rapids' have diminished greatly for *stryk* in 2012-2019, they note that the sense is still in use, and that the largest actual change in the sense of *stryk* is likely the addition of the 'fail' sense, rather than the absence of the 'rapids' sense. Overall, there seems to be a tendency for certain scientific senses to be more frequent in the earlier time periods. Examples of this is the 'geological horizon' sense for *horisont*, the botanical sense of 'umbrella' for *skjerm*, and perhaps 'distribution board' for *tavle*.

A common problem for semantic change is the tendency in Norwegian, as in Swedish and German, among other languages, to spell compound nouns as single tokens, without spaces. However, this makes exact matches with compounds containing these words impossible. In English, one might be able to match the word *computer* in the compound *computer game*, while this would not be possible in Norwegian. For example, according to Leksikografisk Bokmålskorpus (LBK)[9] (Ruth Vatvedt Fjeld and Anders Nøklestad and Kristin Hagen, 2020), the word *kode* 'code' occurs word-initially in around 132 lemmas, and word-finally in about 138 lemmas. In some cases the changes associated with a lemma might be more or less visible in only certain compounds, even if the lemma itself has not lost its sense. It could possibly also be that a compound lemma is preferred to the more ambiguous lemma by itself. We do not claim any direct connection between the sense of a noun and the frequencies of its compounds, but believe they can be an indicator of usage. Some examples of this are discussed below.

6.1. Subset 1

Of the 37 words in ('clean') Subset 1, 11 words showed binary semantic change. Details are shown in table 5. This period had 3 more words with binary change than Subset 2, resulting in a higher percentage of binary change. The average graded change is also higher. The words with binary change were *anfektelse* 'distraction', *bit* 'bite', *bølge* 'wave', *forhold* 'relation, relationship', *horisont* 'horizon', *leilighet* 'opportunity, apartment', *mål* 'goal, measure', *pære* 'pear, bulb', *plattform* 'platform', *rev* 'fox', *skjerm* 'screen'. An additional 10 words had graded change score above zero. We will discuss three interesting cases from this period below.

Subset	Words	Binary	Percent.	Average
Subset 1	37	11	29.7	0.26
Subset 2	37	9	24.3	0.22

Table 5: Frequencies for subset 1 and subset 2, indicating the number of binary changes, the corresponding percentage of changed words, and the average graded change.
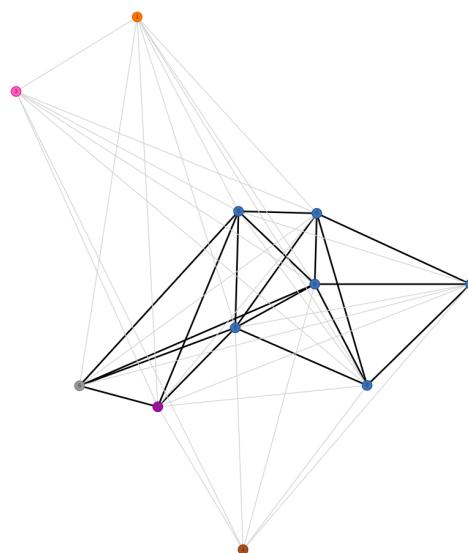

Figure 3: Sense clusters for *bølge* 'wave' from the first time period (1929-1965) The blue cluster marked 0 represents the '(sea) wave' sense, the remaining senses represent metaphorical usage.

Plattform This word had three senses in the earlier time periods, and two in the latter. In the first period, the sense of 'generic platform' dominated with 5 cases, whereas 4 cases were of 'tram platform' and 1 was *oil platform*. Not surprisingly, it is the 'oil platform' sense that dominates in the latter period. It is also interesting to see the disappearance of the 'tram platform' sense, which is likely due to changes in how the tram worked. This change seems typical of the period.

Rev Intended for its possible sense of 'joint', the word *rev* turned out to show other changes over time. Both time periods show frequent usage of the expected sense 'fox', but we observe that the sense of 'reef' has become more frequent.

Pære The sense of 'light bulb' seems to have been present in both periods, but with only 2 cases in the earlier period, and 9 in the latter. Interestingly, the 'pear (fruit)' sense becomes less frequent, according to our data. This is contrary to our original expectations, as one would expect there to be more mention of the electric bulbs. According to the N-gram service, the less ambiguous *lyspære* 'light bulb' is present in both time periods, but steadily increasing, with a higher frequency in the latter period.

6.2. Subset 2

Of the 37 words in ('clean') Subset 2, 8 words showed binary semantic change, while 9 more showed graded change above zero. The remaining 23 words showed no change. The nine words with binary change were *stryk* 'rapids, fail', *kanal* 'channel', *kode* 'code', *oppvarming* 'heating', *innstilling* 'setting', *tavle* '(black)board', *fane* 'banner', *strøm* 'current, electricity'. As with subset 1, we will look at four interesting words from this subset, including the 'questionable' word *mus* 'mouse'.

Stryk was perhaps somewhat surprising. The only word with a perfect graded change score of 1, its score comes from the disappearance of the frequent sense 'rapids', and the introduction of the 'fail' sense. The annotators note that the 'rapids' sense still exists, but the 'fail' sense is new. The national library N-gram service (NB N-gram) shows that indeed, the relative frequency of the word *strykkarakter* 'fail grade', chosen for its unambiguity while being etymologically related to *stryk*, went drastically up towards the end of the 1990's, keeping a higher frequency after its peak.

Kanal This word showed a high degree of change. In the earlier time period, it had 5 senses, where one roughly meaning 'TV and radio channel' became much more frequent in the later time period, while one new sense, seemingly something like 'communication channel', appeared. All other senses disappeared from the word usage graph. Although it is unlikely that the other senses, such as 'river channel' and 'electric channel' have indeed disappeared from language in general, the increase in frequency for the 'TV and radio channel' sense unsurprisingly fits recent trends well.
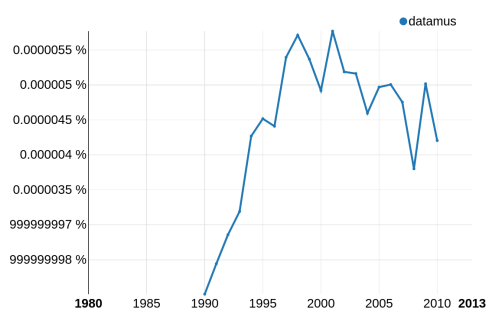

Figure 4: An example of a NB N-gram graph, in this case for the word *datamus* 'computer mouse'. The graph shows that the word appeared in the corpus in 1990 and sharply increased its frequency.

Mus Interestingly, this word lost the sense of 'computer mouse' in the later period, according to our usage graphs. This is the reason why it was judged as 'questionable'. The only sense present in the later time period is that of the animal. However, if we look at the compound *datamus* 'computer mouse', we see that the sense appears in 1990 and becomes more frequent (see figure 4). This strengthens the annotators belief that

the loss of the 'computer mouse' sense might be due to data sampling error.

Spill Somewhat surprisingly, this word did only show a small degree of graded change. Both time periods are dominated by the general sense 'game', and its change is mostly due to the loss of earlier senses. However, if we look at the frequency of the compound noun *dataspill* 'computer game' in the NB N-gram service, we see that although there are a few instances between 1960 and 1980, it is not until the late 1980's that the word becomes more frequent, and it is much more frequent after the 1990's than before.

Fane 'banner, plume' is a good example of how the recent year's digital changes have affected word senses. The word has two senses in the earlier period: 'banner' and 'plume of smoke'. In the newer period, the 'banner' sense prevails, although it seems like its use is somewhat more metaphorical, and in addition we can see the new sense '(website) banner'.

7. Conclusion

We introduced *NorDiaChange*, a dataset of diachronic semantic change on the lexical level for Norwegian. The dataset comprises two new manually annotated subsets that can be used interchangeably as train or test splits. We followed the mainstream DURel framework during annotation, using two historical Norwegian corpora. All the data in the form of diachronic word usage graphs (DWUGs) and accompanying code is available under a CC-BY license at `https://github.com/ltgoslo/nor_dia_change/`. We conducted comprehensive qualitative analysis of the annotation results.

This is the first attempt at developing resources for Norwegian diachronic semantic change, and *NorDiaChange* will become one of the standard NLP benchmarks for Norwegian. In the nearest future, we plan to evaluate various semantic change modelling systems on *NorDiaChange*, and report baseline performances. We are also going to further assess available Norwegian historical text collection in order to come up with a set of reference corpora which are more comparable in terms of their size and genre distribution than the ones we used for *NorDiaChange*.

It is important to note that *NorDiaChange* can be used not only in the field of lexical semantic change detection. By definition, it is also a full-fledged WiC ('word-in-context') dataset (Pilehvar and Camacho-Collados, 2019). As such, it is a ready-to-use benchmark to evaluate *synchronic* word sense disambiguation capabilities of pre-trained language models for Norwegian. Simultaneously, the same models can in theory be used to automatically 'annotate' the existing data, yielding DWUGs for many thousands of words, not just dozens. This is another direction for our future research. Finally, we plan to extend the dataset beyond nouns.

8. Bibliographical References

Blank, A. and Koch, P. (1999). *Historical semantics and cognition*, volume 13. Walter de Gruyter.

Bloomfield, L. (1933). *Language*. Allen & Unwin.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July. Association for Computational Linguistics.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). 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*, pages 2116–2121, Austin, Texas, November. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Florence, Italy.

Kutuzov, A. and Pivovarova, L. (2021a). RuShiftEval: a shared task on semantic shift detection for Russian. In *Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Kutuzov, A. and Pivovarova, L. (2021b). Three-part diachronic semantic change dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online, August. Association for Computational Linguistics.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Pilehvar, M. T. and Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Rodina, J. and Kutuzov, A. (2020). RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

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, June. Association for Computational Linguistics.

Schlechtweg, D., Hätty, A., Del Tredici, M., and Schulte im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy, July. Association for Computational Linguistics.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), December. International Committee for Computational Linguistics.

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, November. Association for Computational Linguistics.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Tahmasebi, N., Borin, L., and Jatowt, A. (2021). *Survey of computational approaches to lexical semantic change detection*. Language Science Press.

Vikør, L. S. (2006). Dei diakrone aspekta ved norsk riksmålsordbok og norsk ordbok. *LexicoNordica*.

9. Language Resource References

Dominik Schlechtweg and Sabine Schulte im Walde. (2021). *Word Usage Graphs (WUGs)*. University of Stuttgart.

Ruth Vatvedt Fjeld and Anders Nøklestad and Kristin Hagen. (2020). *Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk*. Oslo Studies in Language 11(1).