# EMBEDDIA

## Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

# D4.4: Initial cross-lingual news viewpoints identification technology (T4.3)

**Executive summary**

Task T4.3 (Cross-lingual identification of viewpoints and sentiment in news reporting) addresses monolingual and cross-lingual identification of viewpoints and sentiment in news reporting. This deliverable presents initial viewpoints and sentiment detection technology developed. In viewpoints section, we focus on diachronic analysis, more particularly on the changes in word usage across time. We present a range of methods and experiments for semantic change detection in monolingual and multilingual settings. Next, we describe advances in news sentiment analysis, where we first present monolingual neural sentiment analysis methods, followed by cross-lingual experiments, where the languages of training and test sets differ. Last, we present the initial approach on sentiment modelling using subgroup discovery methods.

Partner in charge: JSI

| Project co-funded by the European Commission within Horizon 2020 Dissemination Level | | |
|---|---|---|
| PU | Public | PU |
| PP | Restricted to other programme participants (including the Commission Services) | – |
| RE | Restricted to a group specified by the Consortium (including the Commission Services) | – |
| CO | Confidential, only for members of the Consortium (including the Commission Services) | – |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

## Deliverable Information

| Document administrative information | |
|---|---|
| Project acronym: | **EMBEDDIA** |
| Project number: | **825153** |
| Deliverable number: | **D4.4** |
| Deliverable full title: | **Initial cross-lingual news viewpoints identification technology** |
| Deliverable short title: | **Cross-lingual viewpoints identification** |
| Document identifier: | **EMBEDDIA-D44-CrosslingualViewpointsIdentification-T43-submitted** |
| Lead partner short name: | **JSI** |
| Report version: | **submitted** |
| Report submission date: | **30/06/2020** |
| Dissemination level: | **PU** |
| Nature: | **R = Report** |
| Lead author(s): | **Senja Pollak, Matej Martinc (JSI)** |
| Co-author(s): | **Dragana Miljkovic (JSI), Andraž Pelicon (JSI), Lidia Pivarova (UH), Anita Valmarska (JSI)** |
| Status: | **__ draft, __ final, _x_ submitted** |

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

## Change log

| Date | Version number | Author/Editor | Summary of changes made |
|---|---|---|---|
| 10/05/2020 | v0.1 | Dragana Miljkovic and Senja Pollak (JSI) | Report structure |
| 18/05/2020 | v0.2 | Matej Martinc (JSI) and Lidia Pivovarova (UH) | Diachronic analysis sections |
| 23/05/2020 | v0.3 | Andraz Pelicon (JSI) | Sentiment analysis (SA) classification sections |
| 24/05/2020 | v0.4 | Dragana Miljkovic (JSI) | Croatian dataset description for SA |
| 24/05/2020 | v0.5 | Anita Valmarska (JSI) | Sentiment subgroup discovery |
| 27/05/2020 | v0.6 | Matej Martinc (JSI) | Consolidation of the diachronic analysis section |
| 27/05/2020 | v0.7 | Dragana Miljkovic (JSI) | Edits to the entire report, Sentiment analysis related work |
| 28/05/2020 | v1.0 | Senja Pollak and Dragana Miljkovic (JSI) | final edits for internal review |
| 05/06/2020 | v1.1 | Marko Robnik-Sikonja (UL) | Internal review edits and comments |
| 05/06/2020 | v1.2 | Jose Moreno (ULR) | Internal review edits and comments |
| 15/06/2020 | v1.3 | Matej Martinc, Dragana Miljkovic and Senja Pollak (JSI) | Revised version incorporating reviewers' comments |
| 18/06/2020 | v1.4 | Nada Lavrač (JSI) | Quality control |
| 25/06/2020 | final | Senja Pollak (JSI) | Addressing quality control comments |
| 30/06/2020 | submitted | Tina Anžič (JSI) | Report submitted |

# Table of Contents

# List of abbreviations

| | |
|---|---|
| BERT | Bidirectional Encoder Representations from Transformers |
| CNN | Convolutional Neural Network |
| COHA | Corpus of Historical American |
| DoA | Description of Action |
| EC | European Commission |
| ELMo | Embeddings from Language Models |
| ExM | Ekspress Meedia |
| GA | Grant Agreement |
| JSD | Jensen-Shannon divergence |
| JSI | Jožef Stefan Institute |
| LSTM | Long short-term memory |
| NBM | Naive Bayes Multinominal |
| NE | Named entity |
| NLP | Natural Language Processing |
| OP | Orthogonal Procrustes |
| SA | Sentiment Analysis |
| SD | Subgroup Discovery |
| SemEval | Semantic Evaluation |
| SGNS | Skip-gram negative sampling |
| SVM | Support Vector Machine |
| T | Task |
| TF-IDF | Term Frequency–Inverse Document Frequency |
| WP | Work Package |

# 1 Introduction

The EMBEDDIA project aims to develop monolingual and cross-lingual technology for news media industry. The overall objective of WP4, named *Cross-lingual content analysis*, is to facilitate the analysis of news content across different languages, empowering news media consumers, researchers and news media professionals. This report, entitled *Initial cross-lingual news viewpoints identification technology* is a result of the activities performed in Task T4.3, and focuses developing the technology for identifying viewpoints and sentiment, supporting monolingual, as well as cross-lingual news analysis.

The overall objective of WP4 is to facilitate the analysis of news content in a single, but also across different languages. The specific objectives of WP4 are as follows:

- O4.1 Develop multi-lingual methods for news analysis, aggregation and and topic emergence identification, via Task T4.1
- O4.2 Develop cross-lingual methods to summarize and visualize news, via Task T4.2
- O4.3 Develop methods to identify viewpoints and sentiment in news reporting, via Task T4.3

Specifically, this deliverable at M18 describes on the initial results achieved in viewpoint identification and sentiment analysis (SA) within Task T4.3 (Cross-lingual identification of viewpoints and sentiment in news reporting), which started in M7 and lasts until M33, described in the EMBEDDIA Description of Action (DoA) as follows:

> For the news covering the same events and topics, we will develop methods for detecting viewpoints and sentiments based on media sources. The identification of viewpoints will be done on different levels. For concept level, we will use dynamic word embeddings to analyse contextual changes across time or different news sources. The analyzed news can be either in the same language or, by leveraging WP1's cross-lingual word embeddings, in different languages. On document level, we will employ sentiment level annotation of documents, where cross-lingual embeddings-based models will be tested. Finally, for the news on the same topic (across different sources/countries), the viewpoints will be identified by interpretable classification models and association rules for opinions. The results of this task will be used also in WP3, with the aim of understanding the influence of the news bias/origin on the readership perception. In addition, the gender-representation related analysis will be interpreted in the collaboration with T6.4.

The goal of **viewpoint analysis**, is to reveal differences in coverage of same or similar events. For example, on the concept level analysing of news archives, media partners or media researchers, might be interested in how the word "migrant" changed across time, on the level of news sources one can try to identify the differences in reporting given the political background of newspaper (for example differences between liberal and conservative reporting about legislation of same-sex marriage) or given different cultural background (for example to detect differences in attitudes towards Brexit in UK vs. non-UK press, or a potential topic of interest for our ExM media partners to identify attitudes towards Russia in Estonian and Russian articles of the ExM group).

In this first report, we describe the developed methods using embeddings, especially multilingual contextual embeddings allowing for analysis across different languages. For example, we present a use case on how reporting on *Brexit* changes across time and how the use of word *immigration* in news reporting changes in different time periods and across media sources in different languages. Note that in viewpoints analysis we currently only focus on diachronic aspects - analysing change across time. Since there are no news corpora with manually prescribed semantic change labels, currently the only way to conduct a quantitative evaluation of a system for semantic change detection—without manually labelling a new textual resource—is to conduct experiments on the corpus for which manually labeled evaluation set already exists. Therefore some of the experiments have been conducted on the non-news corpora, focusing on lexical changes. Nevertheless, all the proposed method for semantic change can be leveraged for news analysis and the methods are also directly transferable to detecting non-temporal viewpoints, such as viewpoints in reporting in different countries and newspaper sources.

The other focus of this report is news **sentiment analysis** (SA). SA is one of the most prominent applications of natural language processing (NLP), directly benefiting the companies in analysing public opinions about entities. Another practically important area for media researchers and media companies is—instead of analysing sentiments toward specific target—to analyse the sentiment of a given text itself and to detect how it influences the feelings of the reader, e.g., does a given news provoke positive, negative, neutral sentiment in readers when reading it. From a media researcher's perspective news sentiment is interesting in terms of framing of events and analysing different perspectives, while for media companies, the interest can be in understanding the relation between article's sentiment and its popularity, its appropriateness for placing advertisements or the related behaviour in comments section. In this report we present monolingual neural sentiment analysis methods on news data annotated with sentiment (distinguishing between positive, neutral and negative news), initial cross-lingual experiments, where the languages of training and test sets differ (in particular, we use the available annotated Slovenian news dataset and use it in cross-lingual sentiment classification on a Croatian EMBEDDIA dataset). In addition, we present our initial approach to sentiment modelling using subgroup discovery methods, where we focus on named entities.

The main contributions of T4.3 are as follows:

- **diachronic viewpoints detection technology**: developing methods using embeddings, with the focus on contextual embeddings and aggregation of embeddings through clustering for detecting semantic change of words and allowing for news analysis across time and across languages (work in collaboration of UH and JSI, published in papers Martinc et al. (2020a, 2020b, 2020c); Pivovarova et al. (2019); Marjanen et al. (2019, 2020), see Appendices A–F).
- **new dataset for temporal analysis of English-French cognates**: presents a novel data set and language independent approach to measure comparative semantic change over cognate words (published in Frossard et al. (2020), see Appendix G).
- **monolingual sentiment analysis methods**: on the annotated Slovene news dataset we performed sentiment classification using LSTMs and conducted initial subgroup discovery-based sentiment analysis.
- **initial cross-lingual and multilingual sentiment analysis methods**: using multilingual BERT models, the methods for sentiment classification are trained on available Slovenian sentiment-labelled dataset and tested on EMBEDDIA Croatian news dataset, and the influence of different document representations is analysed.

The deliverable is organised as follows. We start with diachronic news analysis (Section 2), where we provide details on background, methodology and experimental results. The main focus is on several variations of methods for semantic change detection, using static and contextual embeddings in monolingual and multilingual settings. In Section 3 we describe the experiments in monolingual news SA (Sections 3.3), as well as in cross-lingual SA (Section 3.4). The report finishes with the associated outputs of the work done within T4.3, conclusions and plans for further work.

# 2 Diachronic news analysis

Each word has a variety of senses and connotations, constantly evolving through usage in social interactions and changes in cultural and social practices. Identifying and understanding these changes is important for linguistic research and social analysis, since it allows detection of cultural and linguistic trends. It is well-known that some words and phrases can change their meaning completely in a longer period of time. The word *gay*, which was a synonym for cheerful until the 2$^{nd}$ half of the 20$^{th}$ century, is just one of the examples. On the other hand, we are just recently beginning to research and measure more subtle semantic changes that occur in much shorter time periods. These changes reflect a sudden change in language use due to changes in the political and cultural sphere and can, when it comes to media analysis, also reflect a sudden change in a media viewpoint towards a specific entity, practice or person.

Detection of these changes can also be used for improving many Natural Language Processing (NLP) tasks, such as text classification, information retrieval, or word sense disambiguation, which currently to a large extent mistakenly perceive language as a stable and unchanging structure.

In this section, we first present the background and related work on methods for diachronic analysis (Section 2.1), followed by the presentation of several developed methods (Section 2.2) and a range of experiments (Section 2.3), including a monolingual diachronic news analysis on the Brexit dataset, multilingual diachronic analysis applied to news on migrations, analyses of the Corpus of Historical American English (COHA), experiments in the scope of participation in the SemEval 2020 Task 1 shared task, and ideological term analysis on historical press from Finland.

## 2.1 Background and related work

As mentioned in the introduction, for quantitatively evaluating our diachronic viewpoint detection methods, we position our work in the field of lexical semantic change detection, where we can identify two distinct trends: (1) a shift from raw word frequency methods to methods that leverage dense word representations, and (2) a shift from long-term semantic changes (spanning decades or even centuries) to short-term changes spanning years at most (most applicable for news analysis).

Earlier studies (Juola, 2003; Hilpert & Gries, 2008) in detecting semantic and linguistic changes used raw word frequency methods, but are being replaced by methods that leverage dense word representations. The study by Kim et al. (2014) was arguably the first that employed word embeddings, or more specifically, the Continuous Skipgram model proposed by Mikolov et al. (2013). The first research that showed that these methods can outperform frequency based methods by a large margin was conducted by Kulkarni et al. (2015). In the latter method, separate word embedding models are trained for each of the time intervals. Since embedding algorithms are inherently stochastic and the resulting embedding sets are invariant under rotation, vectors from these models are not directly comparable and need to be aligned in a common space (Kutuzov et al., 2018). To solve this problem, Kulkarni et al. (2015) first suggested a simple linear transformation for projecting embeddings into a common space. Zhang et al. (2016) improved this approach by proposing the use of an additional set of nearest neighbour words from different models that could be used as anchors for alignment. Another approach was devised by Eger & Mehler (2017), who proposed second-order embeddings (i.e. embeddings of word similarities) for model alignment. Hamilton et al. (2016a) showed that these two methods can compliment each other.

Since imperfect aligning can negatively affect semantic change detection, the newest methods try to avoid it altogether. Rosenfeld & Erk (2018) presented an approach, where the embedding model is trained on word and time representations, treating the same words in different time periods as different tokens. Another solution to avoid alignment is the incremental model fine-tuning, where the model is first trained on the first time period and saved. The weights of this initial model are used for the initialization of the model trained on the next successive time period. The described step of incremental weight initialization is repeated until the models for all time periods are trained. This procedure was first proposed by Kim et al. (2014) and made more efficient by Peng et al. (2017), who suggested to replace the softmax function for the Continuous bag-of-word and Continuous skipgram models with a more efficient hierarchical softmax. Kaji & Kobayashi (2017) proposed an incremental extension for negative sampling.

Recently, a new type of embeddings called contextual embeddings have been introduced. ELMo (Embeddings from Language Models) by Peters et al. (2018) and BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) are the most prominent representatives of this type of contextual embeddings. In this type of embeddings, a different vector is generated for each context a word appears in. The contextual embeddings solve the problems with word polysemy but have so far not been often used in temporal semantic change studies. The only two temporal semantic change studies we are aware off, that used contextual BERT embeddings, are reported by R. Hu et al. (2019) and Giulianelli (2019).

In the study by R. Hu et al. (2019), contextual BERT embeddings were leveraged to learn a representation for each word sense in a set of polysemic words. Initially, BERT is applied to a diachronic corpus to extract embeddings for tokens that closely match the predefined senses of a specific word. After that, a word sense distribution is computed at each successive time slice. By comparing these distributions, one is able to inspect the evolution of senses for each target word.

In the study by Giulianelli (2019), word meaning is considered as "inherently under determined and contingently modulated in situated language use", meaning that each appearance of a word represents a different word usage. The main idea of the study is to determine how word usages vary through time. First, the authors fine-tune the BERT model on the entire corpus for domain adaptation and after that they perform diachronic fine-tuning, using the incremental training approach proposed by Kim et al. (2014). After that, the word usages for each time period are clustered with the K-means clustering algorithm and the resulting clusters of different word usages are compared in order to determine how much the word usage changes through time.

The second trend in diahronic semantic change research is a slow shift of focus from researching long-term semantic changes spanning decades or even centuries to short-term changes spanning years at most (Kutuzov et al., 2018). For example, a somewhat older research by Sagi et al. (2011) studied differences in the use of English spanning centuries by using the Helsinki corpus (Rissanen et al., 1993). The trend of researching long-term changes continued with Eger & Mehler (2017) and Hamilton et al. (2016b), who both used the Corpus of Historical American (COHA)[1]. In order to test if existing methods could be applied to detect short-term semantic changes in language, newer research focuses on tracing short-term socio-cultural semantic change. Kim et al. (2014) analyzed yearly changes of words in the Google Books Ngram corpus and Kulkarni et al. (2015) analyzed Amazon Movie Reviews, where spans were one year long, and Tweets, where change was measured in months. The most recent exploration of meaning changes over short periods of time that we are aware of, was conducted by Del Tredici et al. (2019), who measured changes of word meaning in online Reddit communities by employing the incremental fine-tuning approach proposed by Kim et al. (2014).

## 2.2 Methodology

In this section we present the core EMBEDDIA methodology for semantic change detection, used for diachronic news analysis. First, we describe how we generate time specific word representations (Section 2.2.1), than we describe how these representations from different time periods can be compared and used for detection of semantic change (Section 2.2.2). We dedicate Section 2.2.3 to the specific case of detection of binary semantic change and Sections 2.2.4 and 2.2.5 to the postprocessing and ensembling techniques, respectively. In Section 2.2.6 we present the evaluation.

### 2.2.1 Word representation

Given a set of corpora containing documents from different time periods, we develop a method for locating words with different meaning in different time periods and for quantifying the meaning changes. For this, we leverage BERT (Bidirectional Encoder Representations from Transformers), a transformer based neural architecture (Vaswani et al., 2017) that employs the transfer learning technique, which has recently become a well established procedure in the field of NLP. This procedure relies on a language model pretraining on very large unlabeled textual resources and after that transfer of the knowledge obtained by the language model onto a specific downstream task by further fine-tuning the model.

In the first step, we fine-tune a pretrained BERT language model for domain adaptation on each corpus. Note that we do not conduct any diachronic fine-tuning, therefore our fine-tuning approach differs from the approach presented in Giulianelli (2019), where BERT contextual embeddings were also used, and also from other approaches from the related work that employ the incremental fine-tuning approach

[1] http://corpus.byu.edu/coha

(Kim et al., 2014; Del Tredici et al., 2019). The reason not to do explicit diachronic fine-tuning lies in the contextual nature of embeddings generated by the BERT model, which are by definition dependent on the context and therefore, in our opinion, do not require diachronic (time-specific) fine-tuning. For all conducted experiments on English corpora, we use the variant of monolingual English BERT model (called BERT-base-uncased) with 12 attention layers and a hidden layer size of 768. For experiments, in which we compare semantic changes across different languages and cultures (see Section 2.3.2), we use the multilingual BERT-base-cased model. For experiments conducted in the scope of the SemEval-2020 Task 1 — Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020) (see Section 2.3.4), we use specific models for each language[2], all with 12 attention layers and a hidden layer of size 768.

After domain adaptation, we generate time specific representations of words. Each corpus is split into predefined time periods and a set of time specific subcorpora is created for each corpus. The documents from each of the time specific subcorpora are split into sequences of byte-pair encoding tokens (Kudo & Richardson, 2018) of a maximum length of 256 tokens and fed into the fine-tuned BERT model. For each of these sequences of length $n$, we create a sequence embedding by summing the last four encoder output layers. The resulting sequence embedding of size $n$ times *embeddings size* represents a concatenation of contextual embeddings for the $n$ tokens in the input sequence. By chopping it into $n$ pieces, we acquire a representation, i.e. a contextual token embedding, for each word usage in the corpus. An entire process of word usage embedding extraction from BERT is visualized in Figure 1. Note that these representations vary according to the context in which the token appears, meaning that the same word has a different representation in each specific context (sequence). After this procedure, we obtain a contextual embedding representation for each word usage, together with the time period each word usage representation belongs to.

Since the byte-pair input encoding scheme (Kudo & Richardson, 2018) employed by the BERT model does not necessarily generate tokens that correspond to words but rather generate tokens that can sometimes correspond to subparts of words, we propose the following *on the fly* reconstruction mechanism that allows us to get word usage representations from byte pair tokens. If a word is split into more than one byte pair token, we take an embedding for each byte pair token constituting a word and build a word embedding by averaging these byte pair tokens. The resulting average is used as a context specific word representation[3].

For some of the experiments in the scope of SemEval-2020 Task 1 (see Section 2.3.4), in addition to contextual embeddings described above, we also generate static word representations by training 300-dimensional Word2Vec model using Skip-gram negative sampling (SGNS) for each time slice. We align embeddings from the different time slices using the Orthogonal Procrustes (OP) method as in Hamilton et al. (2016b). We also applied the pre- and post-processing steps recommended by Schlechtweg et al. (2019).

Similarly, for experiments on the subject of ideological term analysis (see Section 2.3.5), we use the Gensim Word2Vec implementation (Řehůřek & Sojka, 2010) using the Skip-gram model, with a vector dimensionality of 100, window size 5 and a frequency threshold of 100—only lemmas that appear more than 100 times within a double decade are used for training. In this way, we try to ensure that each word in a model has a reliable amount of context and the embeddings are trustworthy. To ensure that embeddings are stable across distinct time slices, we do not conduct aligning, as in the SemEval-2020 Task 1, but rather follow the vector initialization approach proposed by Kim et al. (2014): embeddings for $t+1$ time slice are initialized with vectors built on $t$; then training continues using new data. The learning rate value is set to the end learning rate of the previous model, to prevent models from diverging rapidly.

[2]For English: bert-base-uncased model, for Swedish: bert-base-swedish-uncased (`https://github.com/af-ai-center/SweBERT`), for German: bert-base-german-cased (`https://deepset.ai/german-bert`), for Latin: bert-base-multilingual-uncased model. German is the only language for which we use a cased model since most target words are nouns, which are capitalized in German. The only model available for Latin is a multilingual BERT model trained on 100 languages, including Latin.

[3]This reconstruction mechanism is not employed in the experiments on the Corpus of Historical American English (see Section 2.3.3), where the words that are divided into subparts are simply discarded.

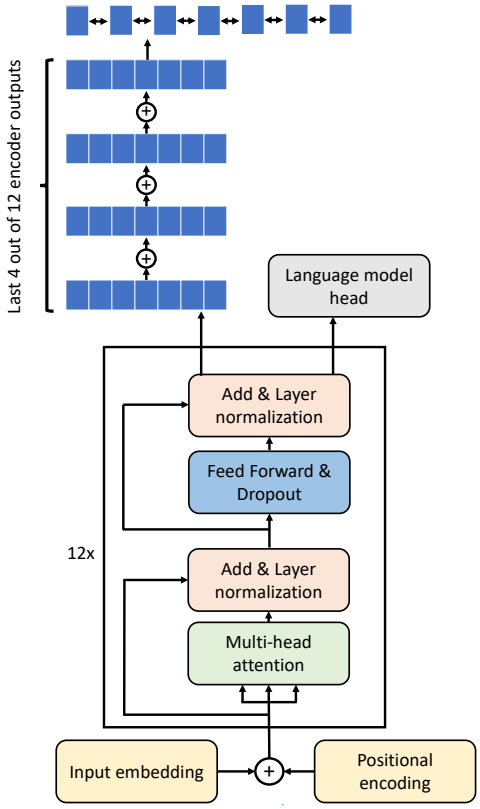



**Figure 1:** Extraction of word usage embeddings from BERT. Note that only the last 4 out of 12 BERT encoder layers are used for the embedding generation.

### 2.2.2 Measures of semantic change

In this section we explain how contextual word usage embeddings can be converted into meaningful time specific representations by two distinct aggregation techniques. We also describe how we measure distance between the aggregated representations in order to asses semantic change.

We employ two methods for aggregating contextual embeddings: averaging and clustering. The first method is more scalable and allows to calculate semantic change for each word in the corpus. The second method (clustering) offers better performance in some settings and is more interpretable, but requires that a set of words, for which the semantic change should be computed, is selected in advance due to scalability limitations.

**Averaging** is a simple aggregation approach where all word usage representations from a given time period are aggregated on the token level (i.e. for every token in the corpus vocabulary, we create a list of all their contextual embeddings) and averaged. A quantitative estimate of semantic change for each word is then measured with the cosine distance between two averaged time-specific representations of the word. Note that for the experiments on the Brexit news corpus (see Section 2.3.1) we conduct the same averaging procedure on the entire corpus (not just on the time specific subcorpus) in order to get a general (not just time specific) representation for each token in the corpus. These general

representations of words are used to find the 50 most similar words to word *Brexit* (see Section 2.3.1 for further details).

**Clustering** of word usage representations results in clusters of word usages, where each cluster is expected to roughly correspond to a single word sense or a specific context. From the output of the clustering algorithms, we create two time-specific cluster distributions by normalizing the cluster counts within each period. We use the Jensen-Shannon divergence (JSD) between two time period-specific distributions to measure the semantic change. The intuition for this approach is the following: if, for instance, a word acquired a novel sense in the latter time period, then a cluster corresponding to this sense only consists of word usages from this period but not the earlier ones, which would be reflected by a higher divergence.

As BERT captures both syntactic and semantic information (Coenen et al., 2019), a cluster does not necessarily correspond to a precise sense of the word but rather represent a specific usage or context (especially in case of analysing the differences in news reporting). Moreover, a word may completely change its context without changing the meaning. Consequently, determining the number of clusters is tricky. Nevertheless, clustering contextual vectors is more interpretable and intuitive than averaging, since it allows to investigate if a word has "gained" or "lost" a specific sense.

For clustering we used the *k-means clustering* algorithm (Steinley, 2006) with various values of $k$ and affinity propagation (Frey & Dueck, 2007)[4]. *Affinity propagation* has been previously used for various linguistic tasks, such as word sense induction (Alagić et al., 2018; Kutuzov et al., 2017). It is based on incremental graph-based algorithm, similar to PageRank. Its main strength is that number of clusters is not defined in advance but inferred during training. We also experimented with the *two-stage clustering* approach inspired by Amrami & Goldberg (2019), where clusters with less than two members are considered weak and merged with the closest strong cluster (see Section 2.3.3 for details).

To obtain a quantitative estimate of semantic change for static embeddings, we measure the cosine distance between the aligned embedding representations of the same word from two time slices.

### 2.2.3 Binary classification of semantic change

By determining the rate of change for a set of target words using one of the semantic change measures described above, these words can be ranked according to amplitude of the measured change. However, a small change indicated by a specific measure does not necessarily indicate that this change is human perceivable or significant. Therefore, in order to determine the significance of changes, in the scope of the SemEval-2020 Task 1, we experimented with two distinct methods for binary classification of semantic change, **thresholding using stopwords** and **identification of period-specific clusters**.

When **thresholding using stopwords** is employed, we try to find the threshold in the ranking list based on the assumption that stopwords—words that are very frequent in a language but do not carry much meaningful information—undergo a low semantic change. We compute semantic change scores for a list of stopwords, employing the same procedure that was used for target words. Next, we compare stopword and target word score distributions in order to define a threshold below which a target word should be classified as unchanged. Full details on thresholding are given in the paper by Martinc et al. (2020c) (see Appendix C).

The second method **identification of period-specific clusters** looks for concrete indications of semantic change, such as appearance or disappearance of a specific word sense. The clustering aggregation techniques cluster target word occurrences in the temporal corpus into a number of distinct clusters. Target word clusters should to some extent resemble different word senses, allowing identification of target words that obtained or lost a meaning. For the most changing words, some clusters may contain mostly word occurrences from one time period. Therefore, if one of the clusters for a target word contains word

[4]We use the Scikit-learn implementations (`https://scikit-learn.org/stable/modules/clustering.html`) with default parameters, except for the number of clusters in k-means algorithm, for which we tried several options.

occurrences from one time period and contains less or equal than $k$ (where k=2) word occurrences from another time period, we assume that this word has lost or gained a specific meaning.

Since clustering methods sometimes produce small-sized clusters, we consider only clusters larger than a threshold, in order to focus on the "main" usages. For k-means clustering we enforce a constraint that a cluster should contain at least 10 word occurrences to be considered in the analysis. For affinity propagation, we implement a dynamic threshold strategy: the threshold for a valid cluster is computed as twice the average cluster size for each target word.

### 2.2.4 Cluster postprocessing

In the scope of the SemEval-2020 Task 1 (see Appendix C), we also tried several heuristics to filter out clusters that potentially contain noise and can distort comparison between time periods:

- removing smallest clusters (containing only one or two instances) whose appearance in a given time period is not significant
- filtering out sentences where a target word is used as a proper noun (target NE)
- removing clusters with too many named entities (radical NE filtering).

More details on the motivation for filtering and details of the implementation are provided in (Martinc et al., 2020c) in Appendix C.

### 2.2.5 Ensembling

Different approaches for semantic change detection were ensembled by multiplying the semantic change scores produced by different methods for each target word. We experimented with different combinations of averaging, clustering and word2vec based methods in order to test the hypothesis that the synergy between contextualised and static embeddings improves the overall performance. Combinations of models that have too strong correlation (Spearman correlation coefficient above 0.8) were discarded.

### 2.2.6 Evaluation

We evaluated the performance of the proposed approaches for semantic change detection by conducting quantitative and qualitative evaluation.

For **quantitative evaluation**, we leverage a number of publicly available evaluation sets for semantic change detection. First one is a small human-annotated dataset created by Gulordava & Baroni (2011). The dataset consists of 100 words from various frequency ranges, labelled by five annotators according to the level of semantic change between the 1960s and the 1990s. They use a 4-points scale from "0: no change" to "3: significant change", and the inter-rater agreement was 0.51 ($p < 0.01$, average of pairwise Pearson correlations). The most significantly changed words from this dataset are *user* and *domain*; words for which the meaning remained intact, are for example *justice* and *chemistry*. The lexical semantic change score for each word is defined as an average of labels prescribed by the annotators.

Another resource used for quantitative evaluation are four SemEval-2020 Task 1 (Schlechtweg et al., 2020) manually labeled evaluation sets, one for each of the four languages in the competition. Here, senses were manually prescribed to words in the evaluation sets and the word's lexical semantic change score is defined as the difference between its normalized sense frequency distributions for the first time period and the second time period (all SemEval-2020 Task 1 corpora only contain two periods), measured by the Jensen-Shannon Distance (J. Lin, 1991)[5]

[5]The details about the annotation procedure, such as number of evaluators and inter-rater agreement is at the time being unknown, since the SemEval-2020 Task 1 cover paper has not yet been published.

We measure Pearson and Spearman correlation between the lexical semantic change score and the model's semantic change assessment for each of the words in the evaluation set in order to assess the performance of the model (stronger correlation indicates better performance of the model).

In order to evaluate our approach to the binary classification of semantic changes, we used four SemEval-2020 Task 1 evaluation sets with manually prescribed binary labels. According to the manually prescribed senses, the words were first classified as either gaining or as loosing a sense, if the sense is attested at most k times in the annotation sample from one time period, but attested at least n times in the sample from another time period. The value of k was set to 0 and n to 1 for Latin; k was set to 2 and n to 5 for English, German and Swedish. The word was classified as changed if it gained or lost a sense and these binary labels were used as the gold standard. The classification accuracy was used for measuring the performance of the classification models in the competition.

**Qualitative evaluation** was used for the experiments on the Brexit news and Immigration news corpora (see Sections 2.3.1 and 2.3.2 for details), since manually labeled news evaluation sets are not available and we were therefore not able to quantitatively assess the approach's performance on these two corpora. The performance of our approach on these two corpora is evaluated indirectly, by measuring how does a specific word of interest semantically correlate to other seed words in a specific time period and how does this correlation vary through time. The cosine distance between the time specific representation of a word of interest and the specific seed word is used as a measure of semantic relatedness. We can evaluate the performance of the model in a qualitative way by exploring if detected differences in semantic relatedness (i.e. relative semantic changes) are in line with the occurrences of relevant events which affected the news reporting about Brexit and Immigration, and also the findings from the academic studies on these topics. This is possible because topics of Brexit and Immigration have been extensively covered in the news and several qualitative analyses on the subject have been conducted.

The argumentation for this type of evaluation comes from structural linguistics and states that a word meaning is a relational concept and that words obtain meaning only in relation to their neighbours (Matthews & Matthews, 2001). According to this hypothesis, the change in the word's meaning is therefore expressed by the change in semantic relatedness to other neighbouring words. Neighbouring seed words to which we compare the word of interest for the Brexit news corpus are selected automatically (see Section 2.3.1 for details) while for the Immigration news corpus, the chosen seed words are concepts representing most common aspects of the discourse about immigration (see Section 2.3.2 for details).

## 2.3 Experiments

In this section we report on the semantic change detection experiments conducted on a number of corpora and languages in several distinct studies, out of which several have already been published (Martinc et al., 2020a, 2020b; Pivovarova et al., 2019; Marjanen et al., 2019, 2020), and the study Martinc et al. (2020c) was submitted (all papers are provided as Appendices).

### 2.3.1 Brexit news

In the first study (Martinc et al., 2020a), **Brexit news corpus** was compiled to test the ability of the proposed approach for semantic change detection to detect relative semantic changes (i.e. how does a specific word, in this case *Brexit*, semantically correlate to other words in different time periods) and to test the method on consecutive yearly periods. The subject of Brexit was chosen due to its extensive news coverage over a longer period of time, which allows us to detect possible correlations between the actual events that occurred in relation to this topic and semantic changes detected by the model. The corpus contains about 36.6 million tokens and consists of news articles (more specifically, their titles and content) about Brexit[6] from the RSS feeds of the following news media outlets: Daily Mail,

[6]Only articles that contain word *Brexit* in the title were used in the corpus creation.

BBC, Mirror, Telegraph, Independent, Guardian, Express, Metro, Times, Standard and Daily Star and the Sun. The corpus is divided into 5 time periods, the first one covering articles about the Brexit before the referendum that occurred on June 23, 2016. The articles published after the referendum are split into 4 yearly periods. The yearly splits are made on June 24 each year and the most recent time period contains only articles from June 24, 2019 until August 23, 2019. The corpus is imbalanced, with time periods of 2016 and 2018 containing much more articles than other splits due to more intensive news reporting. See Table 1 for details.

**Table 1:** Brexit and Immigration news corpora statistics.

| Corpus | Time period | Num. tokens (in millions) |
|---|---|---|
| Brexit news | 2011 - 23.6.2016 | 2.6 |
| Brexit news | 24.6.2016 - 23.6.2017 | 10.3 |
| Brexit news | 24.6.2017 - 23.6.2018 | 6.2 |
| Brexit news | 24.6.2018 - 23.6.2019 | 12.7 |
| Brexit news | 24.6.2019 - 23.8.2019 | 2.4 |
| Brexit news | Entire corpus | 36.6 |

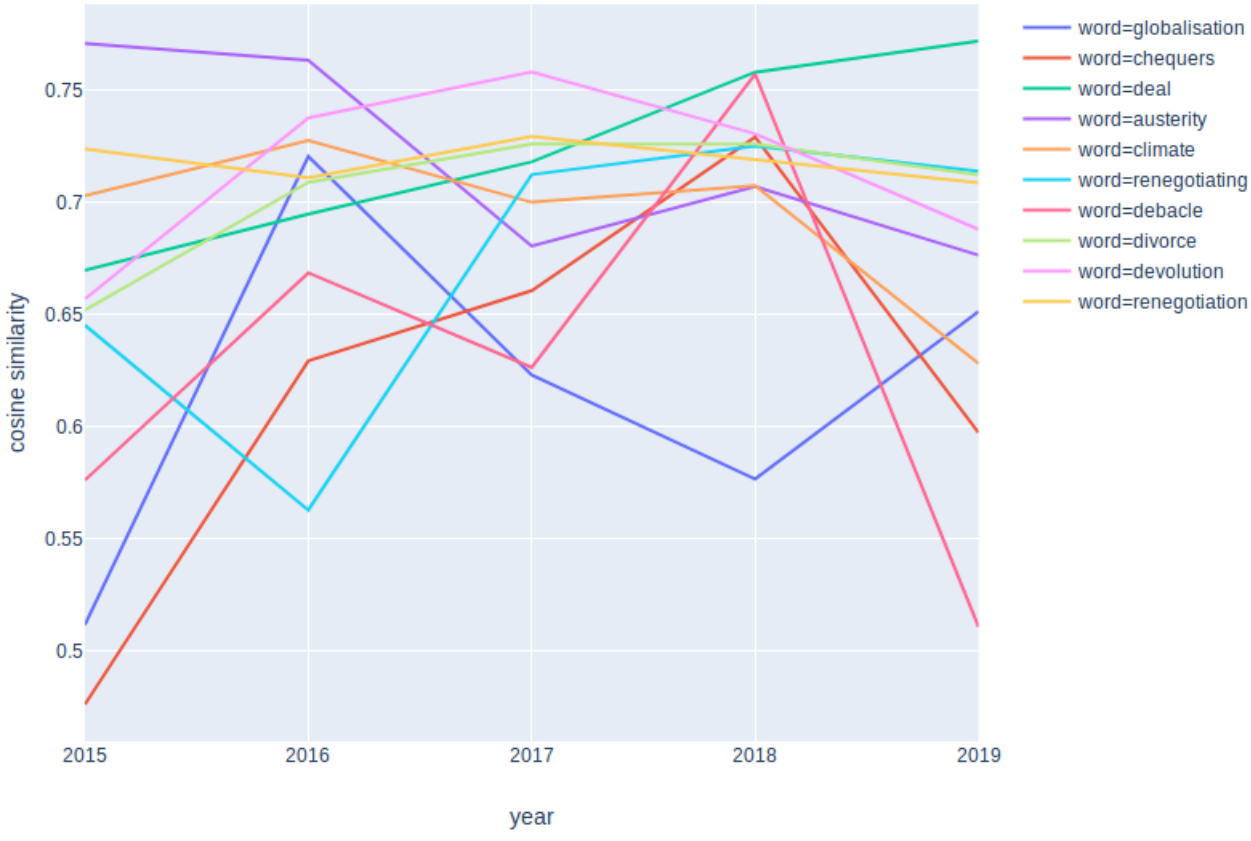



**Figure 2:** The relative diachronic semantic change of the word *Brexit* in relation to the ten words that changed most out of 50 closest words to *Brexit* according to the cosine similarity.

On the Brexit corpus, we employ the **averaging** approach to aggregate contextual embeddings into time period specific representations (see Section 2.2.2) and try to asses the performance of the proposed approach for detecting sequential semantic change of words in short-term yearly periods. More specifically, we explore how time specific seed word representations in different time periods change their semantic relatedness to the time specific word representation of the word *Brexit*. The following proce-

dure is conducted. First, we find 50 words most semantically related to the general non-time specific representation of *Brexit* according to their non-time specific general representations. Since the initial experiments showed that many of the 50 most similar words are in fact derivatives of the word *Brexit* (e.g., *brexitday*, *brexiters*, etc.) and therefore irrelevant for the purpose of this study (as their meaning fully depends on the concept from which they were derived), we first conducted an additional filtering according to the normalized Levenshtein distance defined as:

$$normLD = 1 - (LD/max(len(w1), len(w2))),$$

where $normLD$ stands for normalized Levenshtein difference, $LD$ for Levenshtein difference, $w1$ is *Brexit*, and $w2$ are other words in the corpus. Words for which normalized Levenshtein difference is larger than 0.5 are discarded and out of the remaining words we extract 50 words most semantically related to *Brexit* according to the cosine similarity.

Out of the 50 most similar words, we found ten words that changed the most in relation to the time specific representation of the word *Brexit* with the following equation:

$$MC = abs(CS(w1_{2015}, w2_{2015}) - CS(w1_{2019}, w2_{2019}))$$

where *MC* stands for meaning change, *CS* stands for cosine similarity, $w1_{year}$ is a year specific representation of the word *Brexit* and $w2_{year}$ are year specific representations of words related to *Brexit*.

The resulting 10 seed words are used to determine the relative diachronic semantic change of the word *Brexit* as explained in Section 2.2.6. Figure 2 shows the results of the experiments. We can see that the word *deal* is becoming more and more related to *Brexit*, from having a cosine similarity to the word *Brexit* of 0.67 in 2015 to having a cosine similarity of 0.77 in 2019. This is consistent with the expectations. The largest overall difference of about 0.14 in semantic relatedness can be observed for the word *globalisation*, which was not very related to *Brexit* before the referendum in year 2016 (with the cosine similarity of about 0.52) and than became strongly related to the word *Brexit* in a year after the referendum (with cosine similarity of 0.72). We can observe another drop in similarity in the following two years and then once again a rise in similarity in 2019. This movement could be at least partially explained by the post-referendum debate on whether UK's *Leave* vote could be seen as a vote against globalisation (Coyle, 2016).

A sudden rise in semantic relatedness between words *Brexit* and *devolution* in years 2016 and 2017 could be explained by the still relevant question of how UK's withdrawal from the EU will affect its structures of power and administration (Hazell & Renwick, 2016). We can also observe a sudden drop in semantic relatedness between the words *Brexit* and *austerity* in year 2017, one year after the referendum. It is possible, that the debate on whether UK's *leave* vote was caused by austerity-induced welfare reforms proposed by the UK government in 2010 (Fetzer, 2019) has been calming down. Another interesting thing to note is the extreme drop of about 0.25 in cosine similarity for the word *debacle* after June 23 2019, which has gained the most in terms of semantic relatedness to the word *Brexit* in 2018. It is possible that this gain is related to the constant delays in the UK's attempts to leave the EU.

Some findings of the model are harder to explain, such as differences in *renegotiating* and *renegotiation* curves and changes in relations of *Brexit* to *chequers* and *climate*. For more detailed discussion see our paper by Martinc et al. (2020a) in Appendix A.

### 2.3.2 Immigration news

**Immigration news corpus** was compiled to test the ability of the proposed approach to detect relative semantic changes in a multilingual setting, something that has to our knowledge never been tried before. The main idea is to detect similarities and differences in semantic changes related to immigration in two distinct countries with different attitudes and historical experiences about this subject.

**Table 2:** Immigration news corpus statistics.

| Corpus | Time period | Num. tokens (in millions) |
|---|---|---|
| Immigration news | 2015 | 2.2 |
| Immigration news | 2016 | 2.6 |
| Immigration news | 2017 | 2.6 |
| Immigration news | 2018 | 2.6 |
| Immigration news | 2019 | 1.9 |
| Immigration news | Entire corpus | 11.9 |

The topic of immigration was chosen due to relevance of this topic for media outlets in both countries that were covered, England and Slovenia. The corpus consists of 6,247 English articles and 10,089 Slovenian news articles (more specifically, their titles and content) about immigration[7]; it is balanced in terms of the number of tokens for each language and altogether contains about 12 million tokens. The English and Slovenian documents are combined and shuffled[8] and after that the corpus is divided into 5 yearly periods (split on December 31). The English news articles were gathered from the RSS feeds of the same news media outlets as the news about Brexit, while the Slovenian news articles were gathered from the RSS feeds of the following Slovenian news media outlets: Slovenske novice, 24ur, Dnevnik, Zurnal24, Vecer, Finance and Delo.

We asses the performance of the proposed averaging aggregation approach in a multilingual English-Slovenian setting. Since the main point of these experiments is to detect differences and similarities in relative semantic changes in two distinct languages, we first define English-Slovenian word pairs that arguably represent some of the most common aspects of the discourse about immigration (Martinez Jr & Lee, 2000; Borjas, 1995; Heckmann & Schnapper, 2016; Cornelius & Rosenblum, 2005). These English-Slovenian matching translations are *crime-kriminal*, *economy-gospodarstvo*, *integration-integracija* and *politics-politika*. We measure the cosine similarity between time specific vector representations of each word in the word pair and a time specific vector representation of a word *immigration*.

The results of the experiments are presented in Figure 3. We can note that in most cases and in most years English and Slovenian parts of a word pair have very similar semantic correlations to the word *immigration*, which suggest that the discourse about immigration is similar in both countries. The similarity is most apparent for the word pair *crime-kriminal* and to a slightly lesser extent for the word pair *politics-politika*. On the other hand, not much similarity in relation to the word *immigration* can be observed for the English and Slovenian words for economy. This could be partially explained with the fact that Slovenia is usually not a final destination for modern day immigrants (who therefore do not have any economical impact on the country) and serves more as a transitional country (Garb, 2018), therefore the immigration is less likely to be discussed from the economical perspective.

Figure 3 also shows some interesting language specific yearly meaning changes. The first one is the rise in semantic relatedness between the word *immigration* and the English word *politics* in 2016. This could perhaps be related to the Brexit referendum which occurred in the middle of the year 2016 and the topic of *immigration* was discussed by politicians from both sides of the political spectrum extensively in the referendum campaign.

Another interesting yet currently unexplainable yearly change concerns Slovenian and English words for *integration* in 2019. While there is a distinct fall in semantic relatedness between words *integration* and *immigration*, we can observe a distinct rise in semantic relatedness between words *integracija* and *immigration*. More details about the experiment can be found in Martinc et al. (2020a), attached as Appendix A.

[7] The corpus contains English articles that contain words *immigration*, *immigrant* or *immigrants* in the title and Slovenian articles that contain Slovenian translations of these words in either title or content.

[8] Shuffling is performed to avoid the scenario where all English documents would be at the beginning of the corpus and all Slovenian documents at the end, which would negatively affect the language model fine-tuning.

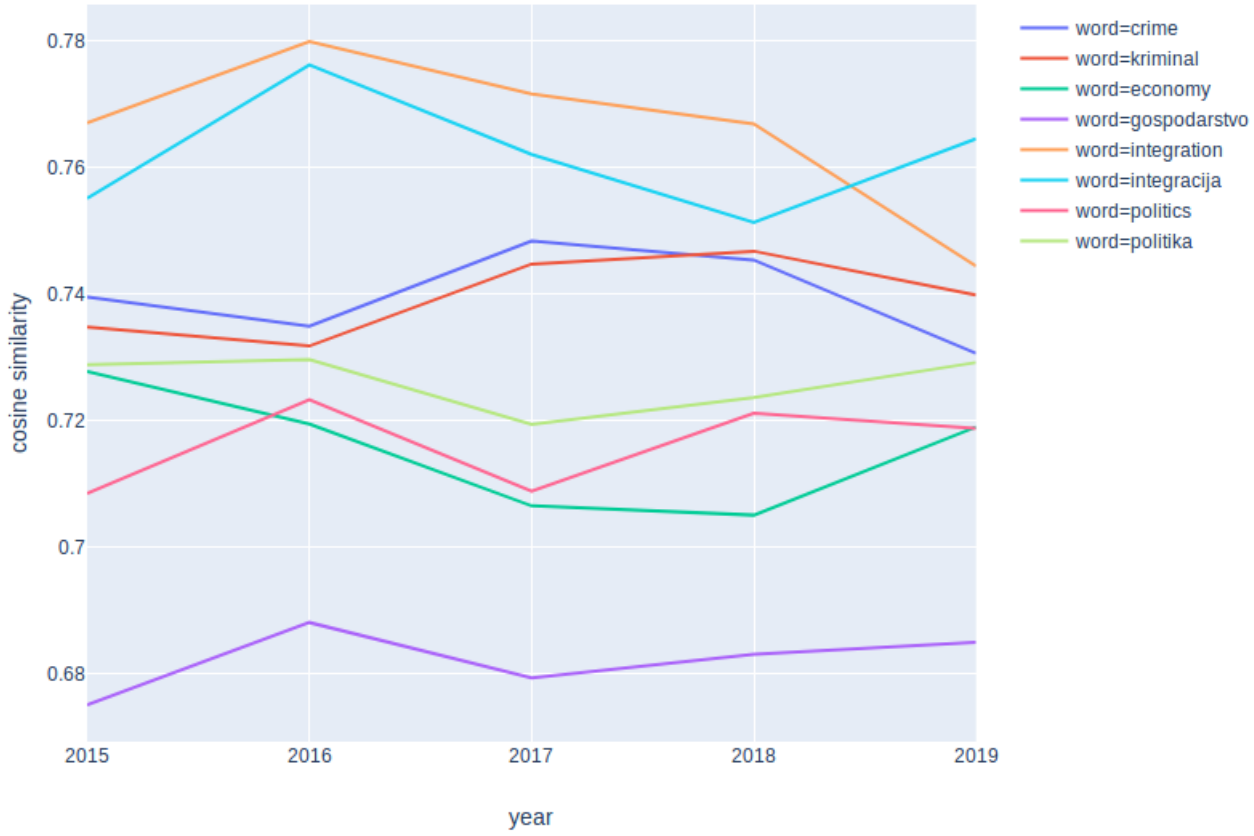



**Figure 3:** The relative diachronic semantic change of the word *immigration* in relation to English-Slovenian word pairs crime-kriminal, economy-gospodarstvo, integration-integracija and politics-politika.

### 2.3.3 Corpus of Historical American English

Since there are no news corpora with manually prescribed semantic change labels, currently the only way to conduct a quantitative evaluation of a system for semantic change detection without manually labelling a new textual resource, is to conduct experiments on the corpus, for which manually labeled evaluation set already exists. One of the most commonly used evaluation sets is the one created by Gulordava & Baroni (2011), which contains manually prescribed semantic shifts for a set of target words between 1960s and 1990s (see Section 2.2.6). To evaluate our approach on this test set, we fine-tuned the BERT model on part of the Corpus of Historical American English (COHA)[9]. The entire corpus contains more than 400 million words from the 1810s-2000s, where data from each decade are balanced by genre—fiction, magazines, newspapers, and non-fiction texts, gathered from various Web sources. In our study (Martinc et al., 2020b, in Appendix B), we focus on the most recent data in this corpus, from the 1960s to the 1990s (1960s has around 28 million and 1990s has 33 million words), to match the manually annotated data. We fine-tuned the BERT model only on this subset.

In the experiments, the focus is on comparing various clustering approaches and metrics for detection of semantic changes. Table 3 shows the Pearson and Spearman correlations between models' outputs and human-annotated drifts. We also report Silhouette scores for clusterings.

A specificity of BERT is the representation of words with byte-pair encodings (Kudo & Richardson, 2018). Thus, some words are divided into several sub-parts; for example, in our list of hundred target words for evaluation, *sulphate* is divided into two byte-pairs *sul* and *##phate*, where *##* denotes the splitting of the

[9] https://www.english-corpora.org/coha/

word. This is also true for the words *medieval*, *extracellular* and *assay*. We decided to exclude these words from our analysis. Thus, strictly speaking, our results are not directly comparable to other approaches in the literature that do not employ BERT.

At the top of Table 3 we overview previous work on the same test set. To train the models, Gulordava & Baroni (2011) used Google Books Ngrams, Frermann & Lapata (2016) used an extended COHA corpus, and both Giulianelli (2019) and Kutuzov (2020) used a subcorpus of COHA, identical to the one used in our experiments. In fact, the setting in Giulianelli (2019) is quite similar to our work, though our best model performance is much higher than in Giulianelli (2019).

As can be seen in Table 3, affinity propagation on the fine-tuned BERT model yields the highest Spearman rank correlation. Results obtained using pretrained and fine-tuned models are consistent: in both runs averaging yields lower performance than clustering, and affinity propagation is the best clustering method. The difference in performance between k-means and affinity propagation could be partially explained by the different number of clusters in the two approaches. Affinity propagation, which performs the best, outputs a huge amount of clusters—160 on average. Two-stage clustering works better than k-means but slightly worse than affinity propagation. Fine-tuning BERT improves all models except k-means with 3 clusters and averaging—we do not yet have a clear explanation for that exception.

Results presented in Table 3 imply that most of the proposed approaches for semantic change detection manage to outperform previous approaches by a large margin. We believe the differences in the numerical results should be primarily attributed to the differences in the methods, even though we can not draw a direct comparison to some of the approaches due to test set word removal and differences in the train corpora.

However, we can compare our results directly to the results published by Giulianelli (2019) since they are also using BERT trained on the COHA corpus. Even more, their proposed clustering approaches are methodologically similar to the approaches presented in this work, yet we manage to outperform

**Table 3:** Correlations between detected semantic changes and manually annotated list of semantic drifts (Gulordava & Baroni, 2011) between 1960s and 1990s.

| Method | Pearson | Spearman | Silhouette |
|---|---|---|---|
| **Related work** | | | |
| **Gulordava & Baroni (2011)** | 0.386 | - | - |
| **Frermann & Lapata (2016)** | - | 0.377 | - |
| **Giulianelli (2019)** | 0.231 | 0.293 | - |
| **Kutuzov (2020)** | 0.233 | 0.285 | - |
| **Pretrained BERT** | | | |
| **Averaging** | **0.354** | **0.349** | - |
| **k-means, k = 3** | 0.461 | 0.444 | 0.104 |
| **k-means, k = 5** | 0.476 | 0.443 | 0.096 |
| **k-means, k = 7** | 0.485 | 0.434 | 0.091 |
| **k-means, k = 10** | 0.478 | 0.443 | 0.086 |
| **2-stage clustering, Aff. propagation** | 0.530 | 0.485 | - |
| **Affinity propagation** | **0.548** | **0.486** | 0.039 |
| **Fine-tuned BERT for 5 epochs** | | | |
| **Averaging** | 0.317 | 0.341 | - |
| **k-means, k=3** | 0.411 | 0.392 | **0.105** |
| **k-means, k=5** | 0.539 | 0.508 | 0.098 |
| **k-means, k=7** | 0.526 | 0.491 | 0.092 |
| **k-means, k=10** | 0.500 | 0.466 | 0.088 |
| **k-means, k=100** | 0.315 | 0.337 | 0.042 |
| **2-stage clustering, Aff. propagation** | 0.554 | 0.502 | - |
| **Affinity propagation** | **0.560** | **0.510** | 0.043 |

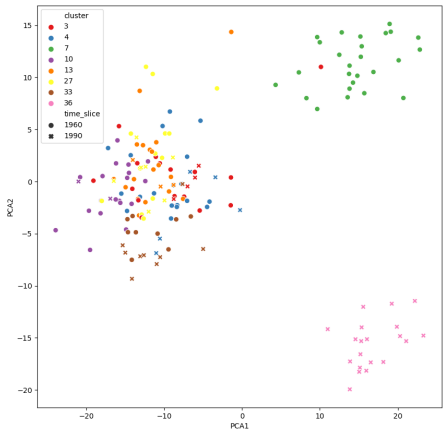



**Figure 4:** 2D PCA visualization for the biggest clusters obtained for word *neutron*.

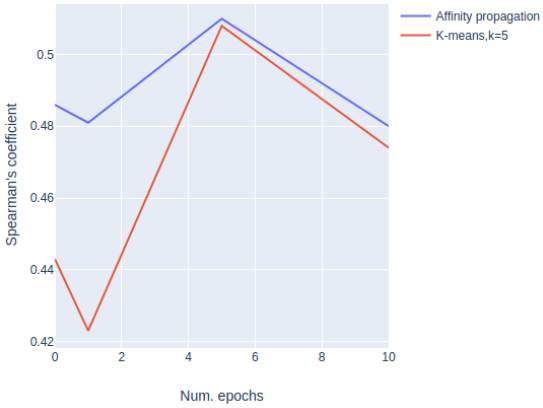



**Figure 5:** Impact of BERT fine-tuning on the performance of two distinct aggregation methods, affinity propagation and k-means with k=5.

their approach by a margin of about 35 percentage points when affinity propagation is used and by about 33 percentage points with k-means clustering[10], the same as in Giulianelli (2019).

Unfortunately, Giulianelli (2019) does not report a number of clusters that has been used, they only mention that the number of clusters has been optimised using the Silhouette scores. We can only speculate why their results are much lower than ours. The first hypothesis is connected with the usage of the Silhouette score, which might not be optimal for our goals. We computed the Silhouette score[11] for clusterings obtained by our methods. As can be seen in Table 3, the best Spearman correlation coefficient does not correspond to the best Silhouette score. Moreover, the Silhouette scores are close to zero.

The second hypothesis for the cause of differences is connected with the difference in fine-tuning regimes employed in this research and the one conducted by Giulianelli (2019). We use domain adaptation fine-tuning, proving its efficiency for a certain number of epochs, for both k-means (except for a small number of clusters) and affinity propagation. Giulianelli (2019) tried both diachronic fine-tuning (using the incremental fine-tuning technique first proposed by Kim et al. (2014)) and domain-specific fine-tuning, but concluded that none led to an improvement in the results. As it was already speculated in Giulianelli (2019), using both training regimes at the same time might lead to too extensive fine-tuning and therefore over-fitting. In future, a more thorough study on influence of incremental fine-tuning on contextual embeddings models (such as BERT) should be conducted, since the effects might differ from the ones observed for static embeddings models.

For the purposes of **error analysis**, we manually checked a few examples by choosing the words that have less mentions in the corpus to be able to look through all sentences containing these words. One of tricky cases for our model is the word *neutron*: according to the manual annotation, it is ranked 81st and has a stable meaning, while our best model considered it to be among the most changed and ranked it at 9.

We visualize the biggest clusters for *neutron* using PCA decomposition of BERT embeddings (Figure 4). There are two clearly distinctive clusters: cluster 36 in the bottom right corner, drawn with pink crosses, which consists only of instances from 1990s, and cluster 7 drawn with green dots in the top right corner, which consists only of instances from 1960s. A manual check reveals that the former cluster consists of sentences which mention *neutron stars*. Though neutron stars have been already discovered in 1960s they were probably less known[12] and are not represented in the corpus. In any case, a difference

[10] Here we are referring to our best k-means configuration with five clusters and using the BERT model fine-tuned for five epochs.

[11] Using standard Scikit-learn implementation, `https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient`

[12] `https://en.wikipedia.org/wiki/Neutron_star`

in the collocation frequency does not mean a semantic change, since collocations often have a non-compositional meaning. Our method could be improved by removing stable multiword expressions and named entities from the training set. The latter distinctive cluster for *neutron*, consisting of word usages from 1960s, contains many sentences that have a certain pathetic style and elevated emotions, such as underlined in the examples below:
*throughout the last several decades the dramatic revelation of this new world of matter has been dominated by a most remarkable subatomic particle – the neutron .*
*the discovery of the neutron by sir james chadwick in 1939. marked a great step forward in understanding the basic nature of matter .*

The lack of such examples in 1990s might have a socio-cultural explanation or it could be a mere corpus artefact. In any case, this has nothing to do with semantic change and demonstrates the ability of BERT to capture other aspects of language, including syntax and pragmatics.

Figure 5 shows the comparison of fine-tuning influence for two best clustering methods (affinity propagation, and k-means with k=5). Interestingly, a light fine-tuning (just for one epoch) decreases the performance of both methods (in terms of Spearman correlation) in comparison to no fine-tuning at all (zero epochs). After that, the length of fine-tuning until up to 5 epochs is linearly correlated with the performance increase. Fine-tuning the model for five epochs appears optimal. The difference between model's performance on 5 epochs is negligible. However, this effect holds only with k=5, other values of k do not demonstrate such a difference between original and fine-tuned models, as can be seen in Table 3. For more details on these experiments, please refer to Martinc et al. (2020b) in Appendix B.

### 2.3.4 SemEval

In order to further evaluate and develop our approach to semantic change detection, we participated in the SemEval-2020 Task 1 — Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020), which deals with detection of semantic change in temporal corpora containing texts from two distinct time periods in four languages: English, German, Latin and Swedish. The statistics about the corpora are presented in Table 4.[13] The challenge defines two subtasks: Subtask 1 is a binary classification, i.e. one has to determine whether a word has changed or not; SubTask 2 aims at ranking a set of target words according to their rate of semantic change.

**Table 4:** SemEval-2020 Task 1 corpora statistics.

| Language | Num. tokens in 1st time period | Num. tokens in 2nd time period |
|---|---|---|
| English | 6.6 millions | 6.8 millions |
| German | 7.0 millions | 7.2 millions |
| Latin | 1.8 millions | 9.4 millions |
| Swedish | 7.1 millions | 11.1 millions |

The results for the binary classification are shown in Table 5. The best official result was achieved by applying the stopword thresholding method (see Section 2.2.3 for details) to rankings obtained by measuring the JSD between affinity propagation cluster distributions (see Section 2.2.2 for details). The method of identifying period-specific clusters worked competitively when conducted on k-means clusters but performed worse with affinity propagation, since the latter method usually produces a large number of clusters. Reducing the number of clusters by merging the closest clusters together increased the efficiency of the method.

The stopword thresholding method seems to work best with higher thresholds, which classify less words as changed. For all methods we face high discrepancies between languages. These can be clearly seen for the method of identifying period-specific clusters deployed on the affinity propagation clustering,

[13] All corpora were given in a lemmatized form and the sentences in the corpora were shuffled, meaning that the contextual information for each word was limited to its sentence.

**Table 5:** SubTask 1 results: accuracy.

| Model | Binary classification method | AVG | English | German | Latin | Swedish |
|---|---|---|---|---|---|---|
| k-means 5 | time-period specific clusters | 0.600 | 0.649 | 0.542 | 0.500 | **0.710** |
| aff-prop | time-period specific clusters, dynamic threshold | 0.496 | 0.568 | 0.458 | **0.700** | 0.258 |
| aff-prop, merging cluster | time-period specific clusters, dynamic threshold | 0.545 | 0.514 | 0.542 | 0.575 | 0.548 |
| aff-prop | stopword tresholding, high threshold | 0.573 | 0.622 | 0.604 | 0.550 | 0.516 |
| aff-prop | stopword tresholding, low threshold | 0.552 | **0.703** | 0.667 | 0.450 | 0.387 |
| ensemble: averaging + aff-prop | stopwords, low threshold | **0.621** | 0.568 | **0.688** | 0.550 | 0.677 |

**Table 6:** SubTask 2 results: Spearman correlation with the ground truth.

| | Input | Method | Post-Processing | AVG | English | German | Latin | Swedish |
|---|---|---|---|---|---|---|---|---|
| *Clustering* | | | | | | | | |
| 1 | pretrained BERT | aff-prop, JSD | - | 0.278 | 0.216 | 0.488 | **0.481** | -0.072 |
| 2 | fine-tuned BERT | aff-prop, JSD | - | 0.298 | 0.313 | 0.436 | 0.467 | -0.026 |
| 3 | fine-tuned BERT | aff-prop, JSD | small clusters | 0.302 | 0.327 | 0.440 | 0.472 | -0.030 |
| 4 | fine-tuned BERT | aff-prop, JSD | target NE | 0.300 | 0.328 | 0.426 | 0.467 | -0.023 |
| 5 | fine-tuned BERT | aff-prop, JSD | NE | 0.295 | **0.436** | 0.302 | 0.467 | -0.025 |
| 6 | fine-tuned BERT | aff-prop, JSD | NE, small clusters | 0.291 | 0.413 | 0.310 | 0.472 | -0.029 |
| 7 | fine-tune BERT | kmeans k=5, JSD | - | **0.320** | 0.189 | **0.528** | 0.324 | **0.238** |
| *Methods not using clustering* | | | | | | | | |
| 8 | fine-tuned BERT | averaging, cosine dist | - | **0.397** | 0.315 | 0.565 | **0.496** | 0.212 |
| 9 | word2vec OP | cosine dist | Schlechtweg et al. (2019) | 0.394 | **0.341** | **0.691** | 0.131 | **0.413** |
| *Ensembling* | | | | | | | | |
| 10 | aff-prop (#2) + w2v (#9) | distance multiplication | - | 0.417 | 0.357 | **0.642** | 0.366 | 0.303 |
| 11 | aff-prop (#2) + w2v (#9) | distance multiplication | NE, small clusters | 0.442 | **0.361** | 0.603 | **0.460** | 0.343 |
| 12 | aff-prop(#2), k-means (#7), averaging (#8), w2v (#9) | non-weighted multiplication | - | 0.403 | 0.279 | 0.607 | 0.451 | 0.276 |
| 13 | aff-prop(#2), k-means (#7), averaging (#8), w2v (#9) | weighted multiplication | - | **0.465** | 0.330 | 0.610 | 0.438 | **0.484** |

which worked the best for Latin and the worst for all the other languages. Overall, the method of identifying period-specific clusters performed better for Swedish and Latin, while stopword thresholding worked better for English and German.

Overall, our team qualified as 11th out of 37 participating teams on Subtask 1.

Results for SubTask 2 are presented in Table 6. The best official result was obtained by the ensemble of word2vec static embeddings and fine-tuned BERT contextual embeddings further improved with radical NE filtering (see Section 2.2.4 for details) as a postprocessing step—see row $\#11$ in the table. The ensemble of four different methods—affinity propagation, K-means (k=5), averaging and word2vec OP—allows merging of all the information that they gather ($\#12$). We can improve this method by taking the correlation between the gold standard and each method as their respective ensembling weights (for each language) ($\#13$). This yields better performance than the best of our submitted methods, though this is not an unsupervised approach and could only be done during the post-evaluation phase.

Ensembles aside, cosine distance between averaged contextual embeddings performs much better than between word2vec representations for Latin but worse for other languages (rows #8 and #9). Affinity propagation clustering, which was the best on the COHA corpus, did not perform well (rows #1 to #6), especially for Swedish, where it performed close to random. One explanation for this discrepancy could be sentence shuffling in the shared task corpora. This means that there is only a limited context available to BERT models during embedding extraction, and BERT models cannot leverage the usual sequence of either 256 or 512 tokens as a context in this setting but are limited to the number of tokens in the sentence. This could have a detrimental effect on the quality of their contextual embeddings. However, the results suggest that by averaging these embeddings, a static embedding of good quality for each target token can be obtained. We note that the lack of content is not problematic for word2vec, which in most cases requires a much smaller contextual window for optimal performance.

Many approaches we tried improved performance only for English: BERT fine-tuning, affinity propagation clustering, and NE filtering. This might be related to the fact that the corpora are lemmatized, and lemmatization has less effect on English, with its reduced morphology. Poor results on the Swedish corpus might be related to OCR-errors, leading to a large number of out-of-vocabulary tokens.

Overall, our team ranked 5th out of 37 teams in Subtask 2, and achieved best results of all teams on Latin language. For more details refer to (Martinc et al., 2020c, Appendix C).

### 2.3.5 Ideological terms analysis

In the work by Pivovarova et al. (2019); Marjanen et al. (2019, 2020) (Appendices E, F and G) we investigate how can some of the computational techniques mentioned in Section 2.2 be leveraged for socio-political research of historical news corpora. This part is done in collaboration with historians on the material of the 19th century press from Finland. We expect that the same method could be applied to more rapid changes in the modern press.

We focus at words with suffix *-ism*. These are terms that help us navigate complex social issues by using a simple one-word label for them. On one hand they are often associated with political ideologies, e.g., *socialism, liberalism*, and on the other they are present in many other domains of language, especially culture, science, and religion, e.g., *impressionism, magnetism, pietism*. Words with the suffix *-ism* are indispensable terms for understanding politics and society, yet they are complex words that give rise to plenty of confusions.

We apply a corpus-based analysis to find out how the vocabulary of isms changed in nineteenth century Finnish newspapers and how usage of ideological isms is different from other words with *-ism* suffix. We try to implement a robust analysis procedure that would be applicable to other tasks with minimal human intervention. Our method consists of two main steps: first, we extract from the corpus *all* words with suffix *-ism*, second, we cluster these words and their semantic neighbours in an unsupervised fashion. This procedure does not require a human intervention other than interpretation of results and, consequently, is potentially applicable to other research questions.

Newspapers in Finland were published in two main languages—Finnish and Swedish. In the beginning of the nineteenth century the majority of newspapers were published in Swedish, though by the 1880s the Finnish and Swedish newspapers were printed in almost equal amount. The Finnish- and Swedish-language press had a different distribution of topics and exposed slightly different political outlook. We use a digitalized collection of nineteenth-century Finnish newspapers freely available from the National Library of Finland (Pääkkönen et al., 2016). We use the full Swedish and Finnish data from 1820 to 1917, treating them as two separate corpora. Each corpus is split into five double-decades. The total amount of words in both corpora is presented in Table 7.

**Table 7:** Corpus size by double decade.

| Time slice | Millions of words | |
|---|---|---|
| | FINNISH | SWEDISH |
| 1820-1839 | 1.3 | 25.5 |
| 1840-1859 | 10.3 | 77.9 |
| 1860-1879 | 90.6 | 326.7 |
| 1880-1899 | 805.3 | 966.9 |
| 1900-1917 | 2439.0 | 953.0 |
| **Total** | 3346.6 | 2355.2 |

To investigate the expansion of the vocabulary of isms, we cluster words into similar groups based on their embeddings using the Affinity Propagation clustering technique (Frey & Dueck, 2007) (see Section 2.2.2). Note that in the experiments described above, the Affinity Propagation was used to cluster contextual embeddings of all usages of a given target word, in order to find its senses. Here, we use affinity propagation to cluster static embeddings of different words to find how they are related.

In the first experiment, we extract from the corpus all ism words. i.e. words that end with *-ism* in Swedish and *-ismi* in Finnish and cluster only this set of words. The extraction allows us to identify how close these words are to each other given other isms in the corpus.

In the second experiment, we try to put isms into a richer context and trace other words associated with them in the respective double-decades. We extract from the corpus all words which have a cosine similarity to any isms that is less than 0.5. We perform clustering on this enriched dataset. Finally, the clusters are filtered so that only the clusters that contain at least one isms word are presented for

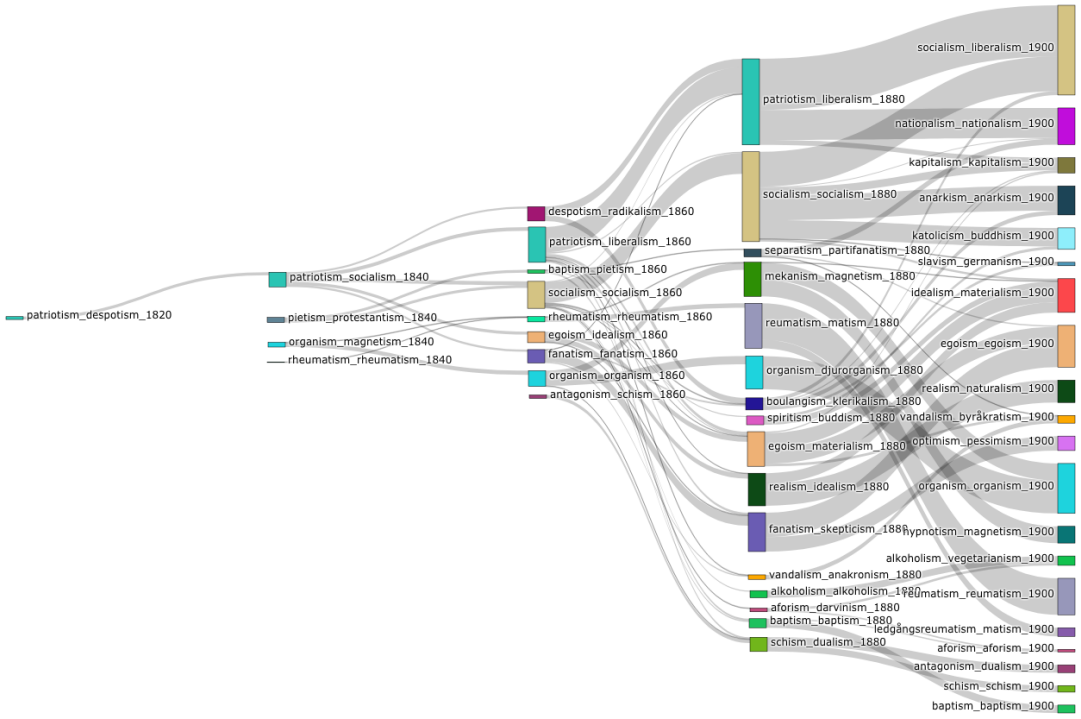



**Figure 6:** Sankey chart of *isms* clusters from the Swedish dataset covering five double decades from 1820 to 1917. The cluster name is the most frequent ism word for that cluster followed by the cluster representative and the double decade.

qualitative analysis. An output of this procedure is different compared to the first experiment, i.e. words that were clustered together in the isms-only clustering, can break up into different enriched clusters, since in the latter setting they have more exemplar options.

Clustering is performed separately for each time slice. To link clusters across time we perform visualization with Sankey charts. In a Sankey diagram, clusters from time slice $t$ are linked to clusters in time slice $t+1$ if they have words in common. The magnitude of the link is the sum of the word frequencies of the common words between the linked clusters from adjacent time slices. We use the frequencies from the source cluster, that is the cluster from time slice $t$.

In Figure 6 we present a Sankey chart for Swedish words ending with *-ism*. Aligning the clusters in the Sankey plots provides a possibility of visually exploring how the vocabulary of isms developed over the course of the century. As can be seen in Figure 6, there is quite a steady expansion of isms from the 1820s onward for Swedish.

One of the main difficulties for our work is a lack of gold standard annotations. We cannot know in advance how the words should be clustered, especially the most problematic ideological terms, which are the main objects of our study. However, we can make several common-sense assumptions on the expected outcome. For example, it would be reasonable to expect that disease names should not appear in the same cluster with philosophical concepts or that artistic movements should be clustered together. This is exactly the case with word *rheumatism*, which is frequently used in the 19th century advertisements.[14]

[14] Automatic advertisement filtering in historical news is not a trivial task since advertisements were less regulated, contained more text and looked similar to other articles.

Table 8, which shows all clusters from our Finnish data that contain words related to rheumatism. It can be seen that *rheumatism* does not interfere with other isms: the clusters consist only from words related to drugs, medical procedures, diseases and other physical conditions, such as baldness or obesity. In that sense, the clusters are rather precise and justify our algorithmic decisions.

**Table 8:** Clusters containing Finnish words related to rheumatism. Original words are presented in italics together with English translations in quotes. *ocr* means the word is incorrectly spelled due to OCR errors; "?" means "impossible to translate"—these are mostly fragments of words appearing due to OCR errors. Bottom left: an advertisement of a rheumatism medicine from *Hufvudstadsbladet*, 01.03.1912, no. 59, p. 15

| 1880-1899 | 1900-1917 |
|---|---|
| *reumatismi* 'rheumatism'<br>*luuvalo* 'gout'<br>*luumalo* 'gout'$_{ocr}$<br>*iskä* '?' *latus* '?'<br>*liikavarvas* 'callus'<br>*kihti* 'gout'<br>*säilöstystauti* 'canning disease'<br>*jalkahiki* 'foot odor'<br>*kivuton* 'painless'<br>*reumatillinen* 'rheumatic'<br>*reumaatillinen* 'rheumatic' | *vähäverisyys* 'anaemia' *risatauti* 'lymphadenitis' *veripuute* 'anaemia' *heillou* 'weakness?'$_{ocr}$<br>*nivelreumatismi* 'arthritis' *epämuodostuma* 'deformity' *kohju* 'hernia'<br>*kroonillinen* 'chronic' *mahatauti* 'gastroenteritis' *mahakatarri* 'gastritis'<br>*suolitauti* 'salt deposits' *riisitauti* 'rickets' *hermovaiva* 'nerve ailment'<br>*verenvähyys* 'anaemia' *ruumisvika* 'body problem' *veritauti* 'blood disease'<br>*lihavuus* 'obesity' *kaljupäisyys* 'boldness' *verettömyydä* 'verettömyydä'<br>*heikkohermoisuus* 'neurasthenia' *lihanen* 'obese' *sukupuoli-* 'sex/gender'$_{ocr}$<br>*sappitauti* 'biliary disease' *heitlous* 'weakness'$_{ocr}$ *selkäydintauti* 'spinal cord disease'<br>*hermoheikkous* 'neurasthenia' *ruokasulatushäiriö* 'digestion problem'<br>*kalvetustauti* 'anaemia' *vinous* 'skewness' *tautitila* 'disease place'<br>*vähäverinen* 'anaemic' *epämuodostua* 'to deform' *hermosairaus* 'neuropathy' |


| |
|---|
| *reumatismi* 'rheumatism' *hiustauti* 'hair disease' *jäsensärky* 'limb ache'<br>*hermo* 'nerve' *oxygeno* '?' *vatsakatar* 'gastritis' *umpitauti* 'constipation'<br>*nuha* 'rhinitis' *hermotautinen* 'neurotic' *topioli* '?' *kurkkukatarri* 'pharyngitis'<br>*parannuskeino* 'remedy' *hoitokeino* 'cure' *spirosiini* 'spirosin' *lazarol* 'lazarol'<br>*lääkitä* 'to medicate' *kotilääke* 'home medicine' *reumaattinen* 'rheumatic'<br>*hammastauti* 'tooth disease' *rautaliuos* 'iron care' *jäsenkolotus* 'limb ache'<br>*leini* 'rheumatism' *linjamentti* 'ointment' *parannusaine* 'betterment' *vilustuminen* 'cold'<br>*luuvalo* 'gout' *latsaro* '?' *hengityselimettauti* 'respiratory disease' |

**Table 9:** Swedish clusters containing word *separatism*

| 1860-1879 | 1880-1899 | 1900-1917 |
|---|---|---|
| *separatism* 'separatism'<br>*mysticism* 'mysticism' *naturalism* 'naturalism'<br>*darwinism* 'darwinism' *moral* 'morality'<br>*tidsanda* 'zeitgeist' *krass* 'crass' *utopi* 'utopia'<br>*materialistisk* 'materialistic' *otro* 'incredible'<br>*rationalistisk* 'rationalistic' *wantro* '?'<br>*menniskonaturen* 'human nature' *tidehvarfvets* '?'<br>*materialism* 'materialism' *materialist* 'materialistic'<br>*konservatism* 'conservatism'<br>*idealism* 'idealism' *rationalism* 'rationalism'<br>*negation* 'negation' *abstraktion* 'abstraction'<br>*idealistisk* 'idealistic' | *separatism* 'separatism' *rent* '?'<br>*finskhet* 'Finnishness' *fennomanins* 'Fennomania'<br>*fennomani* 'Fennomania' *svenskhet* 'Swedishness'<br>*fennomanin* 'Fennomania' *vikingaparti* 'Viking party'<br>*språkpolitik* 'language policy' *publicistisk* 'publishing'<br>*partiagitation* 'party agitation' *partiyra* '?'<br>*partifanatism* 'party fanaticism'<br>*språkgräl* 'language quarrel'<br>*språkfanatism* 'language fanaticism'<br>*språkfråga* 'language question'<br>*språkfrågan* 'language question'<br>*ljusskygghet* 'photophobia' | *separatism* 'separatism' *rikside* 'national idea'$_{ocr}$<br>*statsidé* 'state idea'$_{ocr}$ *rikspolitik* 'national policy'<br>*bourgeoisins* 'bourgeoisie' *byråkratien* 'bureaucracy'<br>*samhällsopinion* 'social opinion'<br>*sträfvandenas* '?' *rikskomplex* 'national complex'<br>*nationalitet-* 'national'$_{ocr}$ *santryska* 'true Russian'<br>*ämbetsmannavälde* 'officialdom'<br>*gränsmärke* 'borderline' *gränsmark* 'borderline'$_{ocr}$<br>*riksenhet* 'national assembly'<br>*samhällskraft* 'social force' *statlighet* 'statehood'<br>*frihetssträvande* 'freedom-aspiring' *wäldets* '?'<br>*riksmakt* 'national power' *självhärskarmakten* '?' |

**Table 10:** Finnish clusters containing word *separatismi*

| 1880-1899 | 1900-1917 |
|---|---|
| *separatismi* 'separatism' *ruotsi-kiihkoinen* 'Svekoman' *ruotsinmielinen* 'Swedish-minded'<br>*ruotsalaisuus* 'Swedishness' *viikinki* 'Viking' *ruotsi-mielinen* 'Swedish-minded'<br>*fennomaani* 'Fennoman' *epäkansallinen* 'anti-national' *viikingit* 'Vikings'<br>*separatisti* 'separatist' *ruotsikko* 'Swedish'(person) *miikinki* 'Viking'$_{ocr}$ *pöppö* '?'<br>*miikingit* 'Vikings'$_{ocr}$ *suomimielinen* 'Finnish-minded' *ruotsi-mielisyys* 'Swedish-mindedness'<br>*wiitinki* 'Viking'$_{ocr}$ *wiilinki* 'Viking'$_{ocr}$ *miitinki* 'Viking'$_{ocr}$ *ruotsimielinen* 'Swedish-minded'<br>*suomi-kiihkoinen* 'Fennoman' *fennoman* 'Fennoman' *henkiheimolainen* 'soul mate'<br>*dagbladilainen* 'member of the Dagblad circle' *miiking* 'Viking'$_{ocr}$ *fennomani* 'Fennoman'<br>*wiiking* 'Viking'$_{ocr}$ *fennomaaninen* 'Fennoman' *ruotsikiihkoisuus* 'Svekomania'<br>*wiilinli* 'Viking'$_{ocr}$ *miikinkilehti* 'Vikings' newspaper'$_{ocr}$ *suomenmielinen* 'Finnish-minded'$_{ocr}$<br>*miikinkiläinen* 'Vikingish'$_{ocr}$ *ruolsinmielinen* 'Swedish-minded' *ruotsiliihloinen* 'Svekoman'$_{ocr}$<br>*herranenluokka* '?' *miikingilehti* 'Vikings' newspaper'$_{ocr}$ *epälansallinen* 'anti-national'$_{ocr}$ | *separatismi* 'separatism'<br>*nationalismi* 'nationalism' *natsionalismi* 'nationalism'<br>*opportunismi* 'opportunism' *natfionalismi* 'nationalism'$_{ocr}$<br>*eristäytyminen* 'isolation' *kansalliskiihko* 'nationalism'<br>*intelligens* 'intelligence' *länsieurooppalainen* 'Western-European'<br>*rotutaistelu* 'race fight' *vapaamielisyy* 'liberalism'$_{ocr}$<br>*sanomalehdistö!* 'press' *antipatia* 'antipathy'<br>*kansallinenviha* 'national anger' *kiihkokansallisuus* 'nationalism'<br>*eristäytyä* 'self-isolate' *liittolaisuus* 'alliance'<br>*vihamieli-syy* 'hostility'$_{ocr}$ *kansallinenylpeys* 'national pride'<br>*kielipolitiikka* 'language policy'<br>*kansallinenliike* 'national movement' |

A more tricky concept is *separatism*, which undergo a noticeable usage change in our datasets as can be seen in Table 9, where we present clusters for Swedish *separatism*. The cluster 1860-1879 presents a religious context of *separatism*, while the 1880-1899 cluster contains completely different set of words, which are related to a contemporary discussion about national identity and national language. The 1900-1917 cluster is again different from the previous two and contains more general political lexis.

The Finnish clusters for *separatismi*, presented in Table 10. The 1880-1899 and 1900-1917 Finnish clusters follow the same pattern as observed for Swedish: the former contains quite specific references, while the latter consists of more general political words. More detailed analysis is provided in the attached papers in Appendices E, F and G.

To sum up, we consider the obtained clusters useful for humanities studies since they provide a researcher with a condensed representation of word usages in a large corpus. Further improvements of the method should include both parts, namely embeddings and clustering. The contextualized embeddings, described in the previous sections of this deliverable, would allow us to investigate gradual semantic shifts rather than split data into discrete time slices, though this would impose scalability issues, since a number of vectors to consider would increase enormously. Improvement of clustering might include fine-tuning of the algorithm parameters, though this is quite hard to do without manually annotated data.In future we aim at finding other applications for the proposed procedure that would be meaningful from newspaper research point of view and easily assessed at the same time.

## 2.4 Conclusion on diachronic analysis and future work

We have presented several methods for diachronic analysis. For majority of experiments, the methods leveraged BERT contextual embeddings in combination with averaging and clustering of embeddings (including various cluster postprocessing techniques), while as baseline approaches we used similar methods using word2vec embeddings. The experiments were performed on a variety of languages, in monolingual and in cross-lingual settings. In future, we will further improve the methods, e.g., in terms of scalability and automated cluster labeling. In terms of applications, we plan to use the developed methods on our newly constructed corpus for temporal for diachronic analysis of English-French cognates (Frossard et al., 2020, Appendix G), on media partners' datasets, as well as for novel applications related to viewpoints analysis, e.g., focusing on difference in perspectives in news media from different ideological background. More detailed plans are presented in concluding section (Section 5).

# 3 Sentiment analysis

Sentiment analysis (SA) is likely the most popular application of NLP, since it can offer a direct benefit to companies, e.g., to summarize customers' opinions. SA has found many areas of applications in customers' product reviews, survey textual responses, social media, etc. It analyzes users' opinions on various topics, such as politics, health, education, etc. A less researched but nevertheless prominent field of research in sentiment analysis is to shift the focus from analyzing sentiment towards a specific target to analyzing the intrinsic mood of the text itself. Several works try to model feelings (either positive, negative or neutral), that readers feel while reading a certain piece of text, especially news (Bučar et al., 2016; Liu, 2012). Van de Kauter et al. (2015) claim that the news production directly affect the stock market as the prevalence of positive news boosts market growth, while the prevalence of negative news impedes it. For this reason, sentiment analysis was applied to the analysis of sentiment in news articles, and especially news articles from the financial domain since researches try to predict the changes in the market based on the sentiment in the news.

In this section, we first explain the background and selected related work (Section 3.1), followed by presentation of the datasets used in our experiments (Section 3.2), including the newly labeled Croatian news dataset. The core of this part of the report are experiments on classification of Slovenian news presented in Section 3.3 (including preliminary experiments in sentiment analysis with subgroup discovery methods in Section 3.3.3), and cross-lingual experiments in Section 3.4.

## 3.1 Background and related work

We present below a brief overview of the state-of-the-art studies relevant for the project, which are related to labelling of datasets for the purpose of sentiment and emotion analysis and methods used for the sentiment analysis.

### 3.1.1 Labelling of datasets for sentiment and emotion analysis

In this section we define emotions as "relatively brief episode of response to the evaluation of an external or internal event as being of major significance" (Scherer, 2000). In our work sentiment is defined as a positive or negative feeling underlying the opinion, which is the definition provided by Liu (2015).

Most common categories in SA are positive and negative sentiment. Sometimes more subtle categories are derived from these two, such as strongly positive, weakly positive, weakly negative and strongly negative. Category neutral appears in many SA studies and is used for news articles from the studied corpora that do not express any sentiment polarity (Bučar et al., 2016).

Labelling of texts for the purpose of emotion analysis is far more complex than for the SA. Emotions categories are in vast majority of studies defined in accordance to Ekman's Ekman (1992) or Plutchik's (Plutchik et al., 1980) emotions categorization. Ekman has defined the following basic emotions: anger, disgust, fear, happiness, sadness and surprise (Ekman, 1992). Plutchik's model has two additional categories compared to Ekman's model: trust and anticipation (Plutchik et al., 1980). There are many lexicons such as SentiWordNet (Baccianella et al., 2010), General Inquirer [15], lexicon by M. Hu & Liu (2004). Additionally, NRC Emotion Lexicon (created by the National Research Council of Canada) contains both SA labelling: positive/negative attitudes and emotions labelled with respect to Plutchik's emotion categories (Mohammad & Turney, 2013). Kadunc & Robnik-Šikonja (2017) have created Slovene sentiment lexicon and evaluated it on news comments.

Emotions analysis is less often applied to news compared to SA. SemEval Task 14 (Strapparava & Mihalcea, 2007) aims at distinguishing between six emotions: anger, disgust, fear, joy, sadness, and surprise. This study reports considerable drop in inter-rater agreement when labelling text with specific emotions instead of just positive/negative attitudes.

Manual annotation of news texts for the purpose of emotion analysis and SA is a time-consuming process. News texts are on average longer than posts on social media, which makes their annotating slower. Moreover, there is an issue of low inter-rater agreements, which results in a dataset with many ambiguous instances and low-accuracy prediction models (Balahur et al., 2010). This is particularly the case when text is annotated with more than three categories (Balahur et al., 2010). However, with more precise instructions, training of the annotators, and several phases in the annotating procedure, it is possible to obtain a decent inter-rater agreement (Bučar et al., 2018).

There is no unique way for annotating text corpus with emotion and/or sentiment information. Texts can be annotated at the several levels of granularity: document (Yessenalina et al., 2010), paragraph (de Arruda et al., 2015), sentence (Farra et al., 2010), and word-level (Engonopoulos et al., 2011).

There are several annotation studies which report annotation of news articles in languages other than English for the purpose of SA. Arruda et al. (2015) have labelled corpus of Brasilian Portuguese in two ways: the target entity (person, who is the main subject in the paragraph) and polarity of the paragraph with respect to the target. The dataset consisted of 131 news texts consisting of 1447 paragraphs. News were selected based on Tweet accounts of main news producers and their popularity. In total four annotators annotated the news articles with positive, negative and neutral categories.

Bučar et al. (2018) annotated news corpora for SA in Slovene (one of the EMBEDDIA languages) on three levels of granularity: sentence, paragraph, and document levels. Between 2 and 6 annotators independently annotated sentiments of a stratified random sample of 10,427 documents from Slovenian

[15] http://www.wjh.harvard.edu/~inquirer/

news portals. The texts were annotated using the five-level Likert scale (1 – very negative, 2 – negative, 3 – neutral, 4 – positive, and 5 – very positive) on all three levels of granularity. We described this dataset in D4.1 and provide a brief description in Section 3.2.

### 3.1.2 Sentiment analysis approaches

There are two common approaches to sentiment analysis. The first is lexicon-based, using a lexicon of weighted words, and machine learning based. Lexicon-based methods use dictionaries with opinion words and match a given set of words in a text to determine polarity. This approach does not need to preprocess the data or train a classifier, which is in contrast to machine learning methods (Islam et al., 2017). Both approaches can be found in news articles sentiment analysis. For our project, most relevant are machine learning, and in particuar neural approaches.

In pre-neural approaches, support vector machines (SVMs) with TF-IDF document representation have been heavily utilized. Usually, these representations are augmented with additional information about sentiment from the sentiment lexicons. K.-Y. Lin et al. (2009) analysed sentiment in Chinese news, with SVM by combining information from the sentiment lexicon and text itself. Li et al. (2014) used a similar method to test the impact of news on stock rates. Kaur & Kaur (2017) used SVM approach to analyse news written in Panjabi, while Bakken et al. (2016) focused on Norvegian political news article and first classified subjective news, and then positive, negative and neutral sentiments. Bučar et al. (2016) trained several models on the corpus of manually annotated Slovene news. The authors used TF-IDF n-grams and they showed that Naïve Bayes Multinomial (NBM) classifier mostly outperforms SVM.

Lately, deep neural networks became popular for the sentiment analysis. Mansar et al. (2017) used convolutional neural networks (CNN). With the help of convolutional layer, they acquired word-level presentations of individual news articles from the learning corpus and combined them with the sentiment score of individual article, which was obtained with a simple, rule-based model. The attributes were used as input to the fully connected CNN. Their model showed the best performance on the SemEval 2017 challenge (Task 5, subtask 2). Moore & Rayson (2017) compared the performance of two models to predict sentiment in news headlines: SVM and bidirectional LSTM recurrent neural network. The neural network achieved 4-=6 % higher accuracy than the SVM model.

## 3.2 Datasets

This subsection presents two datasets that were used for our SA experiments. The first one is SentiNews dataset (Bučar et al., 2018), which is publicly available and contains news articles in Slovene. The second dataset contains news articles in Croatian, which are also labelled for SA at three levels of granularity. The Croatian dataset is used as a testing dataset in a cross-lingual SA study.

### 3.2.1 SentiNews dataset in Slovene

For the sentiment analysis we used SentiNews[16] (Bučar et al., 2018), which is a manually sentiment annotated Slovenian news corpus. The dataset is described in Deliverable D4.1 and we briefly summarize it here. The dataset contains 10,427 news texts from Slovenian news portals (`www.24ur.com`, `www.dnevnik.si`, `www.finance.si`, `www.rtvslo.si`, `www.zurnal24.si`) which were published between 1st of September 2007 and 31st of December 2013. The texts were annotated by 2 to 6 annotators using the five-level Likert scale on three levels of granularity, i.e. on the document, paragraph, and sentence level. The dataset contains information about average sentiment, standard deviation, and sentiment category, which corresponds to the sentiment allocation according to the average sentiment score corresponding to the question "How did you feel after reading this news?".

[16]The dataset is available at `https://www.clarin.si/repository/xmlui/handle/11356/1110`

We used the document-level dataset, with 10,427 news articles (3,261,327 words) and imbalanced distribution of 3,337 (32%) negative, 5,425 (52%) neutral and 1,665 (16%) positive news, where sentiment category corresponds to the sentiment allocation according to the average sentiment score.

The dataset with annotations on document level, was used in monolingual experiments (see Section 3.3), as well as in cross-lingual experiments, where it was used as the training set (see Section 3.4). For more details refer to D4.1. For the experiments in Section 3.3.3, the data set was additionally annotated with named entities, using the RELDI system by Fišer et al. (2018).

### 3.2.2 Croatian sentiment dataset

Annotation of the Croatian language dataset was done in the scope of the EMBEDDIA project task T4.4, but we describe it in this deliverable as we use it in the experiments. The annotation (organised by TRI and JSI) was carried out by 6 annotators who labelled 2,025 Croatian news articles on three levels: document, paragraph and sentence level in order to correspond to the Slovenian dataset. The details of the annotation procedure are available in Deliverable D4.1 "Datasets, benchmarks and evaluation metrics for cross-lingual content analysis", submitted at M9. The annotation procedure fully matches the Slovenian dataset (Bučar et al., 2018) in order to support cross-lingual experiments.

To evaluate the process of annotation, we explored correlation coefficients using various measures of inter-annotator agreement at three levels of granularity, as shown in Table 11. The first three internal consistency estimates of reliability for the scores, shown in Table 11, normally range between 0 and 1. The values closer to 1 indicate more agreement, when compared to the values closer to 0. The Cronbach's alpha values indicate a very good internal consistency at all levels of granularity. Normally, we refer a value greater than 0.8 to a good internal consistency, and above 0.9 to an excellent one (George & Mallery, 2006). The value of Krippendorff's alpha (Krippendorff, 2004) at the document level of granularity implies a fair reliability test, whereas its values at the paragraph level and sentence level are lower. The Fleiss' kappa values illustrate a moderate agreement among the annotators at all levels of granularity. In general, a value between 0.41 and 0.60 implies a moderate agreement, above 0.61 to a substantial agreement, and above 0.81 to an almost perfect agreement (Landis & Koch, 1977). The Kendall's values indicate a fair level of agreement between the annotators at all levels of granularity. Correspondingly, the Pearson and Spearman values range from -1 to 1, where 1 refers to the total positive correlation, 0 to no correlation, and -1 to the total negative correlation. The coefficients show moderate positive agreement among the annotators, but their values are decreasing when applied to the paragraph and the sentence level. Usually, the values above 0.3 refer to weak correlations, above 0.5 to a moderate,and above 0.7 to a strong correlation (Rumsey & Unger, 2015).

In addition, we observed the correlation between the annotators, and found that one of the annotators slightly differs from the rest, which results in overall lower correlations. Despite the clear instructions, the contents of the texts can sometimes be ambiguous, which makes the annotation more difficult. Our results support the claim by O'Hare et al. (2009), that it can be more difficult to accurately annotate sentences (or even phrases). In general, the sentiment scores by different annotators are more consistent at the document level than at the paragraph and sentence level.

The sentiment of an instance is defined as the average of the sentiment scores given by the different annotators (as in the Slovenian news set). An instance was labelled as:

- Negative, if the average of given scores was less than or equal to 2.4,
- Neutral, if the average of given scores was between 2.4 and 3.6,
- Positive, if the average of given scores was greater than or equal to 3.6.

The distribution in resulting dataset is:

- Document level: 15.1% as positive, 21.5% as negative and 63.4% as neutral,
- Paragraph level: 8.9% as positive, 15.7% as negative and 75.4% as neutral,

- Sentence level: 10.0% as positive, 15.4% as negative and 74.6% as neutral.

**Table 11:** Results of dataset annotation: level of inter-rater agreement for document, paragraph and sentence levels.

| | *Doc.level* | | | *Parag.level* | | | *Sent.level* | | |
|---|---|---|---|---|---|---|---|---|---|
| *aC* | 0.927 | | | 0.888 | | | 0.881 | | |
| *aK* | 0.671 | | | 0.565 | | | 0.548 | | |
| *k* | 0.527 | | | 0.489 | | | 0.441 | | |
| | | | | | | | | | |
| | *min* | *max* | *avg* | *min* | *max* | *avg* | *min* | *max* | *avg* |
| *rP* | 0.544 | 0.824 | 0.682 | 0.488 | 0.719 | 0.572 | 0.425 | 0.706 | 0.558 |
| *rS* | 0.557 | 0.762 | 0.669 | 0.474 | 0.702 | 0.548 | 0.42 | 0.696 | 0.54 |
| *W* | 0.508 | 0.73 | 0.625 | 0.449 | 0.656 | 0.513 | 0.389 | 0.649 | 0.504 |

## 3.3 Monolingual sentiment analysis

In this section, we present the work on monolingual system for sentiment analysis in Slovenian news. All our systems were trained using the news sentiment dataset SentiNews (with document level annotations, see Section 3.2). First, we present monolingual classification using LSTMs (Section 3.3.1), followed by transfer learning experiments (Section 3.3.2), and subgroup discovery methods for sentiment (Section 3.3.3).

### 3.3.1 Monolingual neural classification with LSTMs

The dataset with document-level annotations (positive, negative, neutral) was randomly split into training and testing portion in the ratio 80:20.

We trained five neural classification models which share the same architecture. The pooled-LSTM architecture is based on our work described in Pelicon et al. (2019) where it was used for monolingual hate speech detection. The architecture (see Figure 7) is composed of two feature extraction pipelines. The first pipeline takes as an input a tokenized text, embedded into a common vector space. For mapping tokens to embeddings we used 300-dimensional pretrained fastText word embeddings (Joulin et al., 2016). For the purposes of batch training, all the input texts were zero-padded to the same length which was set to 500 tokens. The length was set according to the distribution of news lengths in the corpus so that as much information as possible was retained. The constructed inputs were sent to the LSTM layer with output size of 120, and to the global max-pooling layer applied to all hidden layer outputs from all timesteps. For regularization purposes, we applied dropout to neurons and recurrent connections between timesteps in the LSTM layer. The dropout rate was set to 0.4.

In the second pipeline, the tokenized input texts were weighted using the TF-IDF weighting scheme. We weighted the word n-grams in the range of 1-5 and character n-grams in the range of 1-7. we included only those n-grams that appear more then five times in the corpus. The n-gram weights were sublinearly scaled.

The feature vectors obtained from both pipelines were concatenated and sent to the classification layer which consists of one hidden linear layer and one output layer. The first layer mapped the input vectors to a 150-dimensional output using the Relu activation function. The output layer normalized the output by applying the softmax function.

The trained models mainly differed in text preprocessing steps and enrichment techniques of the original fastText embeddings. The names of the models suggest the data transformation techniques as well as text preprocessing steps used for each particular model. The acronyms LSTM_BOW denote that the model in question used the double data transformation pipeline architecture described above. The

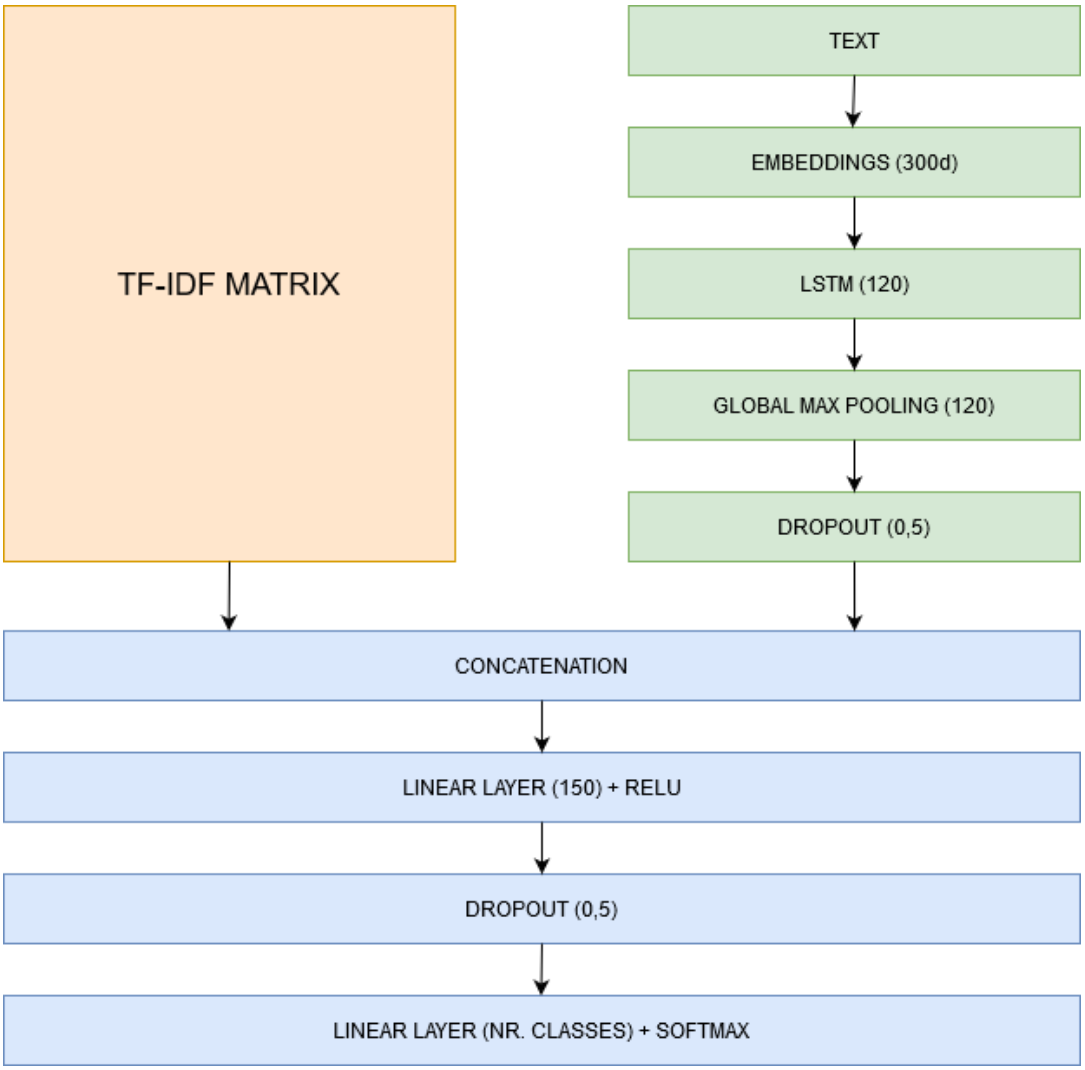



**Figure 7:** Architecture of our monolingual neural classification models.

TPROC acronym denotes the usage of more comprehensive text preprocessing steps. The acronyms REF and REF2 denote models that used sentiment-enriched fastText embeddings.

For the model LSTM_BOW the input texts were only minimally preprocessed, namely only the capitalized characters were lowercased. For the model LSTM_BOW+TPROC, apart from lowercasing the capitalized characters, we also removed all the numbers, stopwords and punctuation from the input texts. Models LSTM_BOW+REF and LSTM_BOW+TPROC+REF utilized the sentiment-enriched fastText embeddings that we produced using a sentiment lexicon based method described in Yu et al. (2017). For embedding enrichment, we utilized the Slovenian sentiment lexicon JOB (Bučar et al., 2016). The models differ in text preprocessing step with LSTM_BOW+REF utilizing lowercased texts as inputs while the preprocessing for the model LSTM_BOW+TPROC+REF utilized the same preprocessing as the model LSTM_BOW+TPROC. The last model LSTM_BOW+REF2 utilized minimal text preprocessing and sentiment-enriched embeddings with additional tuning of the embeddings during the training of the downstream task itself.

All the models were trained for 10 epochs with batch size 32. For optimization, we used the ADAM optimizer with the learning rate of 0.001.

We present the results, i.e. the performance of our models, and compare them to the models from Bučar

**Table 12:** Performance of Slovenian monolingual sentiment classifiers.

| Model | $F_1$ |
|---|---|
| SVM (reported in Bučar et al. (2016)) | 63.4 |
| NBM (reported in Bučar et al. (2016)) | **65.9** |
| SVM (repeated experiment) | 59.0 |
| NBM (repeated experiment) | 24.75 |
| LSTM_BOW | 61.6 |
| LSTM_BOW+TPROC | 61.1 |
| LSTM_BOW+TPROC+REF | 61 |
| LSTM_BOW+REF | 62.1 |
| LSTM_BOW+REF2 | 62.5 |

et al. (2016) where the authors trained SVM and Naive Bayes classifiers on the same dataset. However, the results are not directly comparable as Bučar et al. (2016) evaluated their models with five rounds of ten-fold cross-validation. Due to higher computational complexity of our neural models, we performed less evaluations and used a different split to train-test data. The performance of our neural models, the reimplemented models as well as the reported scores from the original articles are available in Table 12. Since the training corpus was heavily skewed in favor of the neutral class, we benchmarked our models using the macro-averaged $F_1$ score.

The results show that our models outperform the models from the repeated experiments in terms of macro-averaged $F_1$ score. Our weakest models, LSTM_BOW+TPROC and LSTM_BOW+TPROC+REF, perform comparably to the previously reported results of the SVM model in terms of accuracy and lag behind it for approximately 2%. The best performance was achieved by the models with sentiment-enriched fastText embeddings and minimal text preprocessing. It seems that models with more extensive preprocessing lag behind those with minimal preprocessing.

The confusion matrix in Figure 8 shows that our best performing neural model made only a few mistakes between positive and negative classes and suggest that there is a lot of noise in modelling the neutral class.

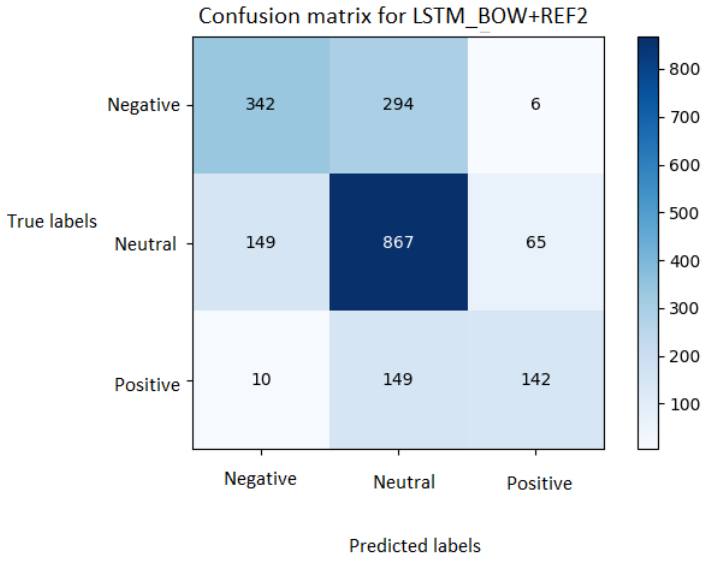



**Figure 8:** Confusion matrix for our best monolingual neural model for sentiment classification of Slovenian news articles LSTM_BOW+REF2.

**Table 13:** Performance of the transfer learning sentiment classifier in comparison with best performing monolingual neural model from section 3.3.1

| Model | F1 |
|---|---|
| LSTM_BOW+REF2 | **0.62** |
| Transfer learning model | **0.62** |

### 3.3.2 Transfer learning experiment

We explored the possibility of boosting the performance of our Slovenian monolingual sentiment classification models through the transfer learning paradigm. We decided to first pretrain the classification model on a related task with a large amount of training data and transfer the resulting model to the actual downstream task of sentiment classification. For the pretraining purpose we chose the task of supervised news categorization. We consider this task to be related to the sentiment classification task as different categories of news tend to contain different distributions of news of different polarity. For example, the crime section of the newspapers and online news platform tend to contain almost exclusively news of negative polarity. As such, we believe that with this task the classifier will obtain relevant information in pretraining that can be leveraged for the downstream task of sentiment classification.

For pretraining our model on news categorization task, we used the corpus of Slovenian news, collected from the online platform of the well-known Slovenian news outlet Dnevnik. All news in the corpus contained the information about the news section in which they appeared on the platform. The pretraining task used information about news sections as weak labels. However, the amount of news in each news section varied considerably. In order to enable our model to learn meaningful patterns, we retained only those news sections which contained at least 200 news. Our final training corpus contained 74,959 news categorized into 50 different categories.

After pretraining, we fine-tuned the model on the same train-test split of the news sentiment dataset as we used in Section 3.2. We fine-tuned all weights of LSTM layers that serve as feature extractors and trained the linear layers that serve for the final classification from scratch. During the initial test runs, we experimented with TF-IDF weigths calculated on the pretraining task, but we decided to recalculate them on the sentiment analysis dataset for the final experiment.

The architecture of our model for this experiment follows the architecture described in the previous section with an additional LSTM layer in one of the feature extraction pipelines. Since pretraining takes a considerable amount of time and proper hyperparameter optimization was infeasible, we initially ran a small number of test runs to estimate the parameters. During test runs we varied the word and character n-gram ranges, number of training epochs and batch size. For the final experiment we utilized the following training parameters:

- Pretraining: Word n-gram range: 1-3, character n-gram range: 1-5, training epochs: 3, batch size: 16
- Finetuning: Word n-gram range: 1-3, character n-gram range: 1-5, training epochs: 10, batch size: 16

During both pretraining and fine-tuning phases we utilized the ADAM optimizer with learning rate 0.001.

We compared the performance of our cross-task transfer learning paradigm with the best neural model from the previous section using macro-averaged $F_1$ score. The results are presented in Table 13. Our transfer learning model achieves comparable performance with the neural model from the previous section, however we did not achieve the boost in performance from the additional pre-training we initially hoped for.

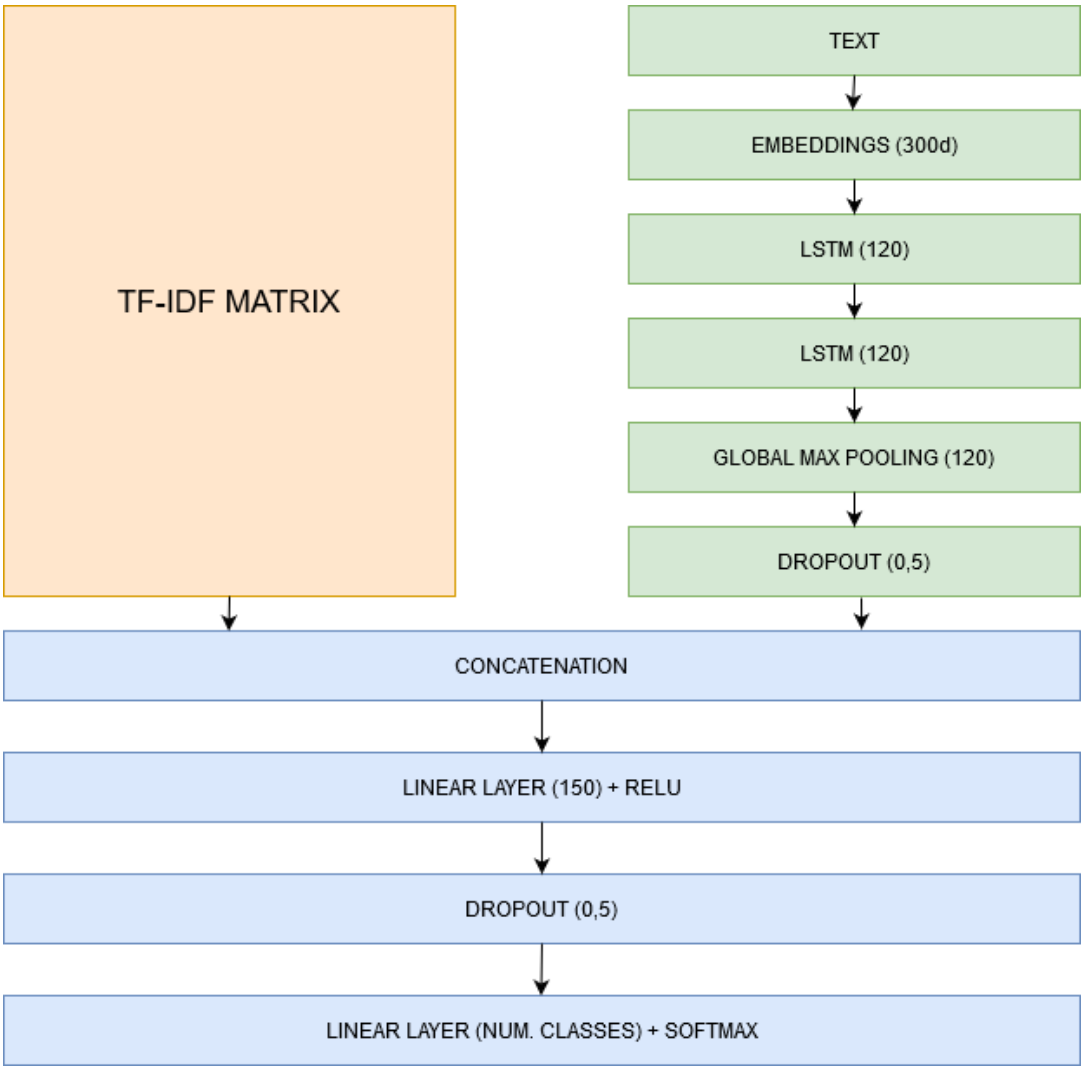



**Figure 9:** Modified architecture for the transfer learning classifier.

### 3.3.3 Sentiment subgroup discovery

In the last set of monolingual experiments, we did not focus on classification, but on subgroup discovery. The goal was to perform exploratory analysis of news sentiment through the lenses of named entities. We are interested in subgroups of articles from the Bučar et al. (2018) dataset, which share named entities and are labelled with either positive or negative sentiment.

Subgroup discovery is a descriptive induction technique that learns descriptive rules from labeled data. The task of subgroup discovery is to find interesting subgroups in the population, i.e. subgroups that have a significantly different class distribution than the entire population (Klösgen, 1996; Wrobel, 1997). The result of subgroup discovery is a set of individual rules, where the rule consequence is a class label. An important characteristic of subgroup discovery is that its task is a combination of predictive and descriptive rule induction. It provides understandable descriptions of subgroups of individuals which share a common target property of interest.

For the purpose of subgroup discovery, we constructed a new tabular dataset, where the attributes are the 10 most frequent named entities[17] from the following categories: person, organization, location, and

[17]The number of most frequent named entities from each category was set arbitrarily, just for the purpose of the experiments.

miscellaneous, thus resulting in 40 attributes. Table 14 presents the list of most frequent named entities from each category. The detection of named entities was done with the ReLDI NER tagger (Fišer et al., 2018) for Slovene. Values presented in Table 14 are the lemmas provided by the ReLDI tagger.

**Table 14:** List of the most frequent named entities from the named entity categories: person, organization, location, and miscellaneous. The quality of identified named entities is low as the identified persons are mostly organizations.

| Person | Organization | Location | Miscellaneous |
|---|---|---|---|
| Dow Jones | Mercator | Slovenija | gorenje |
| Telekom | Petrol | Ljubljana | ZAV |
| Mercatorjev | EU | Evropa | DAX |
| Slovenc | luka KOPER | Krka | Nikkei |
| FTSEurofirst | Krka | ZDA | Nasdaq |
| Gorenjev | nov KBM | nov KBM | Wall Street |
| Moodys | pivovarna Laško | Nemčija | Merkur |
| Telekomov | CAC | Italija | FTSEurofirst 300 |
| Applov | Fed | Triglav | Istrabenz |
| Allianz | NLB | Španija | Sava |

The instances of the newly constructed dataset are constructed from the documents with assigned positive or negative sentiment. The value of the attributes is *yes* if the named entity represented with the attribute has been identified in the document, and *no* otherwise. The class variable is the assigned sentiment of each document. The dataset consists of 4,903 instances—1,616 labelled with positive sentiment and 3,287 with negative sentiment.

We performed subgroup discovery using the DoubleBeam-SD algorithm (Valmarska et al., 2017). The DoubleBeam-SD subgroup discovery algorithm combines separate refinement and selection heuristics with the beam search. Inverted m-estimate is used as the heuristics in the rule refinement phase while m-estimate is the rule selection heuristics. The width of the beam was set to 20. The algorithm shows (at most) 5 best rules.

The rules describing subgroups obtained by the applied subgroup discovery algorithm are presented in Table 15. In the upper half of the table, we present the rules describing subgroups of articles labelled with negative sentiment, while in the bottom half are presented the rules describing subgroups pf articles with positive sentiment. In addition to the rules, we also present the number of covered instances (articles), as well as the statistics of covered true positive and false positive instances. The rules are conjunction of conditions—there are 197 articles which do not mention the organization *Petrol*, and do not mention the location *Krka*, and do not mention *Sava*, and mention the organization *NLB*. Most of covered articles were labelled with the negative sentiment (178), while 19 of the covered articles were labelled with the positive sentiment.

From the results in Table 15 we concluded that there is an interesting subgroup of articles which write about the largest banking group in Slovenia, NLB. The news on the NLB can be associated with strong sentiment, as the bank went through different privatization stages, and polarised Slovenian politics. We performed a post-analysis of the articles where NLB was identified as an organization. In our dataset, there are 210 articles where NLB has been mentioned in the text. Most of these articles are labelled with negative sentiment (185). The articles about NLB were written in the period between September 20, 2007 and November 26, 2013. The articles in our dataset were written by five media outlets in Slovenia: Finance, RTV Slovenia, 24UR, Dnevnik, and Žurnal24. The articles written by the first four outlets were labelled with both positive and negative sentiment. All 9 articles[18] from Žurnal24 were labelled with negative sentiment. Žurnal24 was a free press newspaper (tabloid), and can be considered yellow press.

This will be rectified in future experiments.

[18] Articles were published between May 26, 2009 and May 25, 2013.

**Table 15:** Rules describing subgroups of articles labelled with negative (upper half of the table) and positive sentiment (bottom half of the table).

| Covered instances | TP | FP | Sentiment | | Rule |
|---|---|---|---|---|---|
| 197 | 178 | 19 | negative | ← | org_Petrol = No, loc_Krka = No, misc_Sava = No,org_NLB = Yes |
| 199 | 180 | 19 | negative | ← | loc_Krka = No, misc_Sava = No,org_NLB = Yes, loc_Evropa = No |
| 197 | 178 | 19 | negative | ← | org_luka KOPER = No,org_Petrol = No, loc_Krka = No,org_NLB = Yes |
| 197 | 178 | 19 | negative | ← | org_Petrol = No per_Dow Jones = No, loc_Krka = No,org_NLB = Yes |
| 123 | 68 | 55 | positive | ← | per_Moodys = No loc_Španija = No, per_Slovenc = Yes,org_NLB = No |
| 120 | 66 | 54 | positive | ← | org_pivovarna Laško = No,org_Mercator = No, per_Slovenc = Yes |
| 57 | 34 | 23 | positive | ← | loc_Ljubljana = Yes,org_EU = No, loc_Evropa = Yes |
| 123 | 68 | 55 | positive | ← | loc_Španija = No, per_Slovenc = Yes,org_NLB = No |
| 121 | 67 | 54 | positive | ← | per_Telekom = No, loc_Španija = No, per_Slovenc = Yes, org_NLB = No |

Rules for the target variable *positive* sentiment, mostly cover articles mentioning of *Slovenians* (Slovenc) or *Ljubljana* and *Europe* (Evropa).

In future work, we will perform named entity recognition and named entity linking by utilizing tools developed within WP2 of the project. We expect that this, combined with feature selection, could identify interesting subgroups of articles.

## 3.4 Cross-lingual sentiment analysis

In this section, we present neural network architectures we developed for the cross-lingual sentiment analysis tasks. As explained in Section 3.2, we have a large annotated news dataset for Slovene available. For cross-lingual transfer evaluation, we used the smaller Croatian dataset with the same annotation settings. Note that Slovenian and Croatian languages belong to the same group of Slavic languages.

We developed three different architectures, all based on the multilingual contextual BERT model. The models differ in the way they process the input text documents.

- **Beginning of the document**

  In the first experimental setting, we produced the document representations from only the beginning part of the document. We tokenized the document with the pretrained multilingual BERT tokenizer and took the sequence of 512 tokens from the beginning of the document as an input of the BERT model. We used the [CLS] token vector, produced by the BERT language model, as the document representation and sent it to the linear classification layer.

  For training, we followed the suggestions in the original paper by Devlin et al. (2018). We used the Adam optimizer with the learning rate of $2e-5$ and learning rate warmup over the first 10% of the training instances. For regularization purposes, we used weight decay set to 0.01. We trained the model for three epochs and set the batch size to 16. The batch size was reduced from 32 due to high memory consumption during training that was the result of the long sequence length. In this setting the fine-tuning of the language model and training of the classification layer was performed end-to-end.

- **Beginning and end of the document**

  For the second setting, we constructed document representations from the beginning and end of each document. The length of the input sequence was 512 tokens. After tokenization, we took 256 tokens from the beginning of the text and 256 tokens from the end of the text and concatenated them. We used the sequence as the input of the BERT language model and used the [CLS] token vector from the last layer as the document representation. This document representation was used in the linear classification layer.

The training regime for this model exactly mirrored the training regime for the previous model and is described above.

- **Using sequences from every part of the document**

  In the third setting, we composed document representations from the whole document. We tokenized each document and split it into sequences of 512 tokens. The tokens in each subsequent sequence have the first 50 tokens overlapped with the previous tokens sequence. We attach the same document sentiment label to each of the subsequences from the same document. This oversampled dataset was used to pretrain the multilingual BERT language model with the attached linear layer for classification.

  Once we fine-tuned the multilingual BERT language model, we removed the original classification layer and use the language model as a feature extractor. We prepared each document in the dataset as described above and used every subsequence of a particular document as an input for the BERT model. We extracted the [CLS] vector representations from the last layer and combined them into the final document representation. We tested several ways how to combine the output vector representations into the final document representation.

  - **Using the most informative subsequence representation**

    In this approach, we tried to identify the most informative subsequence for the task at hand. As the BERT language model was fine-tuned on the sentiment classification task, we assumed that some notion of importance of different parts of the text was encoded in the vector representations. Using this reasoning, we defined the most informative subsequence as the subsequence with the highest vector norm. We used only the representation of this subsequence as the final vector representation and discarded the rest. The document representation was sent to a two-layer fully connected neural network which produced the final predictions.

  - **Averaging the representations of all subsequences**

    As the first approach is based on a strong assumption and it does not actually utilize the data from the whole document, we decided to combine all vector representations of subsequences into one final document representation. We simply averaged all the vector representations to produce the final document embedding. The document representation was sent into a two-layer fully connected neural network which produced the final predictions.

  - **Using convolutional layers**

    In this approach, we tried to extract the most informative parts of the document with the use of 1-D convolutional neural layers. We used a convolutional filter of size 2 with stride 2 that runs over the produced subsequence representations. In this way, the convolutional filter processes the subsequences in pairs and extracts the most informative features from each pair of subsequences from each part of the document. The final document representation is sent to the linear layer that produced the final classification.

  The fine-tuning phase of these experiments was the same as the training phase for the previous two models. The classification layers were trained after the fine-tuning phase was completed. Each classification layer was trained for three epochs. We used the Adam optimizer with the learning rate of $2e-5$. For regularization we used weight decay of 0.01.

After initial experiments with different document representation techniques, we tried to boost the performance of the original models by developing a **two-stage system** for sentiment classification. In this approach, we utilized the best performing approach, described above. The two stages of the system were each composed of a separate classifier. The stage one classifier was trained on the whole training dataset to detect articles with neutral sentiment. The stage two classifier was trained to distinguish between articles with positive and negative sentiment. This classifier was trained on a subset of the training dataset, composed only of negative and positive instances. Each of the classifiers was trained

separately on its own subset of the training data, using the same training regime as used for the first two experiments. In this way, the original three-class task that was solved by one model was reduced to two binary classification tasks solved by two separate models. We consider each of these two subtasks as easier than the original task, therefore we expect better individual performances of the two models and consequently better performance of the whole system.

After training each of the two classifiers, the input to the system was fed to the stage one classifier. If this classifier did not classify its input as neutral, the input was fed to the second stage, where the second classifier returned either negative or positive sentiment.

For training the models, we used the dataset by Bučar et al. (2018) described in Section 3.2.1 with 3,337 negative, 5,425 neutral and 1,665 positive news. For testing the cross-lingual performance of our models, we additionally utilized a dataset of Croatian news articles. The dataset contains 2,025 news articles, annotated according to the same scheme as the Slovenian training dataset, as explained in Section 3.2.2.

All the models were first trained and evaluated on the Slovenian training set using 10-fold cross-validation. Performance of the models from each fold was additionally tested on the Croatian test set. The performances from each fold on the Croatian test set were averaged and reported as the final result. The performance of the models was summarized using a standard classification metric, namely macro-averaged $F_1$ score. Due to high imbalance of both training and test set, the accuracy score is not a meaningful metric for this task and it was not reported.

The results of this experiment are presented in Table 16. The performance of our models was compared to the baseline majority classifiers for both Slovenian and Croatian dataset. All models trained in this experiment perform better than the majority baseline classifier by a substantial margin. The best performing model utilizes document representations which include the beginning and end of the original input document. The models that used a combination of every subsequence of the document as a representation performed worse than originally expected. We observed the best performance from the model using the representation that averaged all the subsequence representations. Poor performance was observed when using the 1D convolutional operation to combine all subsequences. However, this method was tested at a disadvantage as only one convolutional filter was used to produce the final document representation.

The best $F_1$ score achieved on Slovenian dataset was 63.6, which is also higher than the performance of our neural monolingual models described in Section 3.3.1. The performance on the Croatian test set drops for 5-10% with the best $F_1$ being 52.9. Interestingly, the lowest drop can be observed with our weakest 1D CNN model, even though the general performance of this model is too low to be of practical use.

After obtaining the performance of each text representation method, we used the best performing model to develop a two stage classification system. The classifiers in each stage of the system utilized the document representation leveraging the beginning and end of the input documents. The results of this

**Table 16:** Performance of our cross-lingual classifiers in cross-validation setting on Slovenian data and on Croatian test set.

| Model | Slovenian cross-validation | Croatian test set |
|---|---|---|
| | F1 | F1 |
| Majority classifier | 22.8 | 26.0 |
| Beginning of the document | 59.5 | 51.2 |
| Beginning and end of the document | **63.6** | **52.9** |
| Sequences from every part of the document | | |
| Most informative subsequence | 58.6 | 49.9 |
| Averaging subsequence representations | 63.3 | 51.1 |
| 1D CNN | 40.7 | 35.1 |

system are reported in Table 17 together with the performance of the majority classifier and the original model that utilizes the same document representation for easier reference.

**Table 17:** Comparison of our two-stage sentiment classifier with our best performing model from the previous experiment in cross-lingual setting.

| Model | Slovenian cross-validation | Croatian test set |
|---|---|---|
| | F1 | F1 |
| Majority classifier | 22.8 | 26.0 |
| Beginning and end of the document | **63.6** | **52.9** |
| Two-stage system | 58.6 | 50.3 |

We may observe that the results of the two-stage system are considerably worse in comparison to the original model even with the most promising document representation technique. Even though the two individual tasks were assumed to be easier to solve, the final system sees a 5% drop in F1 score on the Slovenian dataset as well as a 2.6% drop in F1 score on the Croatian test set.

## 3.5 Conclusions on sentiment analysis and future work

We described advances in news sentiment analysis, where we first presented neural sentiment classification approaches for monolingual and cross-lingual document-level sentiment analysis. The best achieved F1 measure ranges around 63% for monolingual and 52% for zero-shot cross-lingual setting. In future, we will aim at further improving the performance of the systems, and address other datasets, including target-based sentiment and analysis of emotions. We have also performed exploratory analysis using named entities with subgroup discovery methods, which showed promising results, and will be continued with advanced named-entity recognition and linking technologies from WP2. In addition, in collaboration with T4.2 we have performed preliminary visualisation-based sentiment analysis, which is reported in Deliverable D4.3. Conclusions and future plans are in more detail described in Section 5.

# 4 Associated outputs

The work described in this deliverable has resulted in the following resources:

| Description | URL | Availability |
|---|---|---|
| Martinc et al. (2020a) | `https://github.com/EMBEDDIA/semantic_shift_detection` | Public (MIT) |
| Martinc et al. (2020b) | `https://github.com/EMBEDDIA/AddMoreClusters` | Public (MIT) |
| Martinc et al. (2020c) | `https://github.com/EMBEDDIA/Semeval2020-Task1` | Public (MIT) |

Parts of this work are also described in detail in the following publications, which are attached to this deliverable as appendices:

| Citation | Status | Appendix |
|---|---|---|
| Matej Martinc, Petra Kralj Novak, and Senja Pollak. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC). Marseille, France, May 2020, pp. 4813–4821. | Published | Appendix A |
| Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. Capturing Evolution in Word Usage: Just Add More Clusters? In Companion Proceedings of the Web Conference 2020. 2020, pp. 343—349. | Published | Appendix B |
| Matej Martinc, Syrielle Montariol, Lidia Pivovarova, and Elaine Zosa "Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings not Always Better Than Static for Semantic Change Detection. Submitted to International Workshop on Semantic Evaluation (SemEval 2020), colocated with COLING 2020. | Submitted | Appendix C |
| Lidia Pivovarova, Jani Marjanen, and Elaine Zosa. Word Clustering for Historical Newspapers Analysis. In Proceeding of Ranlp Workshop on Language technology for Digital Humanities. 2019: 3. | Published | Appendix D |
| Jani Marjanen, Lidia Pivovarova, Elaine Zosa, and Jussi Kurunmäki. Clustering ideological terms in historical newspaper data with diachronic word embeddings. 5th International Workshop on Computational History, HistoInformatics 2019. CEUR-WS, 2019. | Published | Appendix E |
| Jani Marjanen, Jussi Kurunmäki, Lidia Pivovarova, and Elaine Zosa. The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections. Journal of Data Mining and Digital Humanities. 2020. | Submitted | Appendix F |
| Esteban Frossard, Mickael Coustaty, Antoine Doucet, Adam Jatowt, and Simon Hengchen. Dataset for Temporal Analysis of English-French Cognates. Journal of Data Mining and Digital Humanities. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC). Marseille, France, May 2020, pp. 855–859. | Published | Appendix G |

# 5 Conclusions and further work

In this report we presented initial EMBEDDIA experiments on viewpoint and sentiment analysis.

With regard to *viewpoints analysis*, we have focused on diachronic analysis, where we have developed several methods. For majority of experiments, the methods leveraged BERT contextual embeddings in combination with averaging and clustering of embeddings, while as baseline approaches we used similar methods using word2vec embeddings. The experiments were performed on variety of languages, in monolingual and in cross-lingual settings. For future work on temporal semantic change detection, we plan to investigate how the clusters found by the methods in this work, can be used to interpret different usages of a word in specific time slices. Some experiments on this subject have already been conducted and initial findings indicate that most of the clusters are interpretable, though some particular meanings can be spread among several clusters. What still needs to be done is to automate the labeling of these clusters, i.e. to automatically tag the specific sense of the word they represent. We also plan to improve the scalability of the clustering method for semantic change detection, since it currently does not allow to generate temporal semantic representations for an entire vocabulary of the temporal corpus.

Our analysis hints that clustering BERT token embeddings for a word does not necessarily lead to sense-specific clusters. This conclusion is on par with Coenen et al. (2019). Indeed, BERT's ability do detect distinct word meanings has limitations and it would be interesting to extract only the semantic

parts of the BERT embeddings to direct our analysis towards word meaning instead of general word usage.

In terms of applications, we plan to use the diachronic analysis methods on media partners' datasets. We have already identified a use case with ExM, where we will investigate how methods can be used for identification of relations between concepts, such as migrations, and different political parties. Next, we would like to test the methods for identifying the differences between newspapers from different ideological origins (e.g., liberal vs. conservative media). Finally, we have started to investigate if the developed methods can be used on biomedical data related to the recent COVID-19 crisis.

In this report, we also describe advances in news *sentiment analysis*, where we first presented monolingual neural sentiment analysis methods, followed by cross-lingual experiments, where the languages of training and testing sets differ. For the monolingual setting, we have used SentiNews, dataset with Slovenian news articles. The results show that the best performance was achieved by the models with sentiment-enriched fastText embeddings and minimal text preprocessing. An interesting observation was that models with extensive text preprocessing lag behind those with less preprocessing. We explored the possibility of boosting the performance of our Slovenian monolingual sentiment classification models through the transfer learning paradigm. Our transfer learning model achieves comparable performance with the monolingual neural models, however we did not achieve the boost in performance from the additional pre-training we initially hoped for. Last, we presented the initial experiments in sentiment modeling using subgroup discovery.

We also worked on the cross-lingual sentiment analysis task and developed three different neural architectures leveraging the multilingual contextual BERT model, The architectures differ in the way they process the input text document. The training dataset was Slovenian news articles, and we tested models on EMBEDDIA developed Croatian sentiment dataset. The best scores were produced using the document representations combining beginnings and ends of documents. The best achieved F1 measure on Slovenian ranges around 63% for monolingual and 52% for zero-shot cross-lingual setting.

In future work, we will consider cross-lingual experiments on other languages and datasets, and continue our work on representation learning for longer documents. We will investigate the role of named entities not only in the subgroup discovery setting, but also in document representation for sentiment analysis. We also plan to work on aspect-based sentiment approaches.

# References


Alagić, D., Šnajder, J., & Padó, S. (2018). Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-second aaai conference on artificial intelligence.*

Amrami, A., & Goldberg, Y. (2019). Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598*.

Arruda, G., Roman, N., & Monteiro, A. (2015, 11). An annotated corpus for sentiment analysis in political news..

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Bakken, P. F., Bratlie, T. A., Marco, C., & Gulla, J. A. (2016, December). Political news sentiment analysis for under-resourced languages. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2989–2996). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from `https://www.aclweb.org/anthology/C16-1281`

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., ... Belyaeva, J. (2010, 01). Sentiment analysis in the news..

Borjas, G. J. (1995). The economic benefits from immigration. *Journal of economic perspectives*, *9*(2), 3–22.

Bučar, J., Žnidaršič, M., & Povh, J. (2018, September). Annotated news corpora and a lexicon for sentiment analysis in slovene. *Lang. Resour. Eval.*, *52*(3), 895–919.

Bučar, J., Povh, J., & Žnidaršič, M. (2016, 03). Sentiment classification of the slovenian news texts. In *Proceedings of the 9th international conference on computer recognition systems cores 2015* (p. 777-787). doi: 10.1007/978-3-319-26227-7_73

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F. B., & Wattenberg, M. (2019). Visualizing and measuring the geometry of bert. In *Neurips.*

Cornelius, W. A., & Rosenblum, M. R. (2005). Immigration and politics. *Annu. Rev. Polit. Sci.*, *8*, 99–119.

Coyle, D. (2016). Brexit and globalisation. *Brexit Beckons: Thinking ahead by leading economists*, 23.

de Arruda, G. D., Roman, N. T., & Monteiro, A. M. (2015, November). An annotated corpus for sentiment analysis in political news. In *Proceedings of the 10th Brazilian symposium in information and human language technology* (pp. 101–110). Natal, Brazil: Sociedade Brasileira de Computação. Retrieved from `https://www.aclweb.org/anthology/W15-5614`

Del Tredici, M., Fernández, R., & Boleda, G. (2019). Short-term meaning shift: a distributional exploration. 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)* (pp. 2069–2075).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eger, S., & Mehler, A. (2017). On the linearity of semantic change: Investigating meaning variation via dynamic graph models. *arXiv preprint arXiv:1704.02497*.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169-200. doi: 10.1080/02699939208411068

Engonopoulos, N., Lazaridou, A., Paliouras, G., & Chandrinos, K. (2011, 01). Els: A word-level method for entity-level sentiment analysis. In (p. 12). doi: 10.1145/1988688.1988703

Farra, N., Challita, E., Abou Assi, R., & Hajj, H. (2010, 12). Sentence-level and document-level sentiment mining for arabic texts. In (p. 1114-1119). doi: 10.1109/ICDMW.2010.95

Fetzer, T. (2019). Did austerity cause brexit? *American Economic Review*, *109*(11), 3849–86.

Fišer, D., Ljubešić, N., & Erjavec, T. (2018). The janes project: language resources and tools for slovene user generated content. *Language Resources and Evaluation*. Retrieved from `https://doi.org/10.1007/s10579-018-9425-z` doi: 10.1007/s10579-018-9425-z

Frermann, L., & Lapata, M. (2016). A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, *4*, 31–45.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science*, *315*(5814), 972–976.

Frossard, E., Coustaty, M., Doucet, A., Jatowt, A., & Hengchen, S. (2020, May). Dataset for temporal analysis of english-french cognates. In *Proceedings of the 12th language resources and evaluation conference* (pp. 855–859). Marseille, France: European Language Resources Association. Retrieved from `https://www.aclweb.org/anthology/2020.lrec-1.107`

Garb, M. (2018). Coping with the refugee and migrant crisis in slovenia: the role of the military. *Defense & Security Analysis*, *34*(1), 3–15.

George, D., & Mallery, P. (2006). *Spss for windows step-by-step: A simple guide and reference, 14.0 update (7th edition)*. USA: Allyn and Bacon, Inc.

Giulianelli, M. (2019). *Lexical semantic change analysis with contextualised word representations*. University of Amsterdam - Institute for logic, Language and computation.

Gulordava, K., & 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* (pp. 67–71).

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing* (Vol. 2016, p. 2116).

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Hazell, R., & Renwick, A. (2016). Brexit: its consequences for devolution and the union. *UCL Constitution Unit Briefing Paper, The Constitution Unit*, *19*, 4.

Heckmann, F., & Schnapper, D. (2016). *The integration of immigrants in european societies: National differences and trends of convergence* (Vol. 7). Walter de Gruyter GmbH & Co KG.

Hilpert, M., & Gries, S. T. (2008). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, *24*(4), 385–401.

Hu, M., & Liu, B. (2004, 08). Mining and summarizing customer reviews. In (p. 168-177). doi: 10.1145/1014052.1014073

Hu, R., Li, S., & Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3899–3908).

Islam, M. U., Ashraf, F., Abir, A., & Mottalib, M. (2017, 12). Polarity detection of online news articles based on sentence structure and dynamic dictionary. In (p. 1-5). doi: 10.1109/ICCITECHN.2017.8281777

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, *37*(1), 77–96.

Kadunc, K., & Robnik-Šikonja, M. (2017). *Slovene sentiment lexicon KSS 1.1.* Retrieved from `http://hdl.handle.net/11356/1097` (Slovenian language resource repository CLARIN.SI)

Kaji, N., & Kobayashi, H. (2017). Incremental skip-gram model with negative sampling. *arXiv preprint arXiv:1704.03956*.

Kaur, G., & Kaur, K. (2017, 12). Sentiment detection from punjabi text using support vector machine. *International Journal of Scientific Research in Computer Science and Engineering*, *5*, 39-46. doi: 10.26438/ijsrcse/v5i6.3946

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal analysis of language through neural language models. *ACL 2014*, 61.

Klösgen, W. (1996). Explora: A multipattern and multistrategy discovery assistant. In *Advances in knowledge discovery and data mining* (pp. 249–271).

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology (second edition)*. Sage Publications.

Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web* (pp. 625–635).

Kutuzov, A. (2020). Diachronic contextualized embeddings and semantic shifts. In *press.*

Kutuzov, A., Kuzmenko, E., & Pivovarova, L. (2017). Clustering of Russian adjective-noun constructions using word embeddings. In *Proceedings of the 6th workshop on balto-slavic natural language processing* (pp. 3–13).

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1384–1397).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1).

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014, 04). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*. doi: 10.1016/j.knosys.2014.04.022

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145–151.

Lin, K.-Y., Yang, C., & Chen, H.-H. (2009, 01). Emotion classification of online news articles from the reader's perspective. In (Vol. 1, p. 220 - 226). doi: 10.1109/WIIAT.2008.197

Liu, B. (2012, May). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. Retrieved from http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016 doi: 10.2200/s00416ed1v01y201204hlt016

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press. doi: 10.1017/CBO9781139084789

Mansar, Y., Gatti, L., Ferradans, S., Guerini, M., & Staiano, J. (2017, August). Fortia-FBK at SemEval-2017 task 5: Bullish or bearish? inferring sentiment towards brands from financial news headlines. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (p. 817-822). Vancouver, Canada: Association for Computational Linguistics. doi: 10.18653/v1/S17-2138

Marjanen, J., Kurunmäki, J., Pivovarova, L., & Zosa, E. (2020). The expansion of isms, 1820–1917: Data-driven analysis of political language in digitized newspaper collections. *Journal of Data Mining and Digital Humanities*.

Marjanen, J., Pivovarova, L., Zosa, E., & Kurunmäki, J. (2019). Clustering ideological terms in historical newspaper data with diachronic word embeddings. In *Histoinformatics2019-the 5th international workshop on computational history.*

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020b). Capturing evolution in word usage: Just add more clusters? In *Companion proceedings of the web conference 2020* (pp. 343–349).

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020c). Discovery team at semeval-2020 task 1: Context-sensitive embeddings not always better than static for semantic change detection. In *International workshop on semantic evaluation (semeval).*

Martinc, M., Novak, P. K., & Pollak, S. (2020a, May). Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th edition of the language resources and evaluation conference (LREC)* (pp. 4813–4821).

Martinez Jr, R., & Lee, M. T. (2000). On immigration and crime. *Criminal justice*, *1*(1), 486–524.

Matthews, P. H., & Matthews, P. (2001). *A short history of structural linguistics*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Nips.*

Mohammad, S., & Turney, P. (2013, 08). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*. doi: 10.1111/j.1467-8640.2012.00460.x

Moore, A., & Rayson, P. (2017, August). Lancaster a at SemEval-2017 task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (p. 581-585). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S17-2095 doi: 10.18653/v1/S17-2095

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., & Smeaton, A. (2009, 11). Topic-dependent sentiment analysis of financial blogs. *TSA 2009 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, 6 November 2009, Hong Kong, China. ISBN 978-1-60558-805-6*, 9-16.

Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., & Mäkelä, E. (2016). Exporting Finnish digitized historical newspaper contents for offline use. *D-Lib Magazine*, *22*(7/8).

Pelicon, A., Martinc, M., & Novak, P. K. (2019). Embeddia at semeval-2019 task 6: Detecting hate with neural network and transfer learning approaches. In *Proceedings of the 13th international workshop on semantic evaluation* (pp. 604–610).

Peng, H., Li, J., Song, Y., & Liu, Y. (2017). Incrementally learning the hierarchical softmax function for neural language models. In *Thirty-first aaai conference on artificial intelligence.*

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pivovarova, L., Marjanen, J., & Zosa, E. (2019). Word clustering for historical newspapers analysis. In *Ranlp workshop on language technology for digital humanities.*

Plutchik, R., Kellerman, H., & Press, A. (1980). *Theories of emotion*. Academic Press. Retrieved from `https://books.google.si/books?id=TV99AAAAMAAJ`

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rissanen, M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., ... Raumolin-Brunberg, H. (1993). The helsinki corpus of english texts. *Kyttö et. al*, 73–81.

Rosenfeld, A., & Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 474–484).

Rumsey, D. J., & Unger, D. (2015). *U can: Statistics for dummies*. John Wiley.

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, *73*, 161–183.

Scherer, K. (2000, 01). Psychological models of emotion. *The Neuropsychology of Emotion*.

Schlechtweg, D., Hätty, A., Del Tredici, M., & im Walde, S. 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* (pp. 732–746).

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). Semeval-2020 task 1: Unsupervised lexical semantic change detection. *To appear in SemEval@COLING2020*.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1–34.

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th international workshop on semantic evaluations* (p. 70–74). USA: Association for Computational Linguistics.

Valmarska, A., Lavrač, N., Fürnkranz, J., & Robnik-Šikonja, M. (2017). Refinement and selection heuristics in subgroup discovery and classification rule learning. *Expert Systems with Applications*, *81*, 147–162. doi: 10.1016/j.eswa.2017.03.041

Van de Kauter, M., Breesch, D., & Hoste, V. (2015, July). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Syst. Appl.*, *42*(11), 4999–5010. Retrieved from `https://doi.org/10.1016/j.eswa.2015.02.007` doi: 10.1016/j.eswa.2015.02.007

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of the first european symposium on principles of data mining and knowledge discovery, PKDD 1997* (pp. 78–87).

Yessenalina, A., Yue, Y., & Cardie, C. (2010, October). Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1046–1056). Cambridge, MA: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/D10-1102`

Yu, L.-C., Wang, J., Lai, K. R., & Zhang, X. (2017). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 534–539).

Zhang, Y., Jatowt, A., Bhowmick, S. S., & Tanaka, K. (2016). The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, *28*(10), 2793–2807.

# Appendix A: Leveraging contextual embeddings for detecting diachronic semantic shift

## Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift

**Matej Martinc, Petra Kralj Novak, Senja Pollak**
Jozef Stefan Institute
Ljubljana, Slovenia
{matej.martinc, petra.kralj.novak, senja.pollak}@ijs.si

**Abstract**

We propose a new method that leverages contextual embeddings for the task of diachronic semantic shift detection by generating time specific word representations from BERT embeddings. The results of our experiments in the domain specific LiverpoolFC corpus suggest that the proposed method has performance comparable to the current state-of-the-art without requiring any time consuming domain adaptation on large corpora. The results on the newly created Brexit news corpus suggest that the method can be successfully used for the detection of a short-term yearly semantic shift. And lastly, the model also shows promising results in a multilingual settings, where the task was to detect differences and similarities between diachronic semantic shifts in different languages.

**Keywords:** Contextual embeddings, Diachronic semantic shift, Diachronic news analysis

### 1. Introduction

While language is many times mistakenly perceived as a stable, unchanging structure, it is in fact constantly evolving and adapting to the needs of its users. It is a well researched fact that some words and phrases can change their meaning completely in a longer period of time. The word *gay*, which was a synonym for cheerful until the 2$^{nd}$ half of the 20$^{th}$ century, is just one of the examples found in the literature. On the other hand, we are just recently beginning to research and measure more subtle semantic changes that occur in much shorter time periods. These changes reflect a sudden change in language use due to changes in the political and cultural sphere or due to the localization of language use in somewhat closed communities.

The study of how word meanings change in time has a long tradition (Bloomfield, 1933) but it has only recently saw a surge in popularity and quantity of research due to recent advances in modelling semantic relations with word embeddings (Mikolov et al., 2013) and increased availability of textual resources. The current state-of-the-art in modelling semantic relations are contextual embeddings (Devlin et al., 2018; Peters et al., 2018), where the idea is to generate a different vector for each context a word appears in, i.e., for each specific word occurrence. This solves the problems with word polysemy and employing this type of embeddings has managed to improve the state-of-the-art on a number of natural language understanding tasks. However, contextual embeddings have not yet been widely employed in the discovery of diachronic semantic shifts.

In this study, we present a novel method that relies on contextual embeddings to generate time specific word representations that can be leveraged for the purpose of diachronic semantic shift detection [1]. We also show that the proposed approach has the following advantages over existing state-of-the-art methods:

- It shows comparable performance to the previous state-of-the-art in detecting a short-term semantic shift without requiring any time consuming domain adaptation on a very large corpus that was employed in previous studies.
- It enables the detection and comparison of semantic shifts in a multilingual setting, which is something that has never been automatically done before and will facilitate the research of differences and similarities of how word meanings change in different languages and cultures.

The paper is structured as follows. We address the related work on diachronic semantic shift detection in Section 2. We describe the methodology and corpora used in our research in Section 3. The conducted experiments and results are presented in Section 4. Conclusions and directions for further work are presented in Section 5.

### 2. Related Work

If we take a look at a research on diachronic semantic shift, we can identify two distinct trends: (1) a shift from raw word frequency methods to methods that leverage dense word representations, and (2) a shift from long-term semantic shifts (spanning decades or even centuries) to short-term shifts spanning years at most.

Earlier studies (Juola, 2003; Hilpert and Gries, 2008) in detecting semantic shift and linguistic change used raw word frequency methods for detecting semantic shift and linguistic change. They are being replaced by methods that leverage dense word representations. The study by Kim et al. (2014) was arguably the first that employed word embeddings, or more specifically, the Continuous Skipgram model proposed by Mikolov et al. (2013), while the first research to show that these methods can outperform frequency based methods by a large margin was conducted by Kulkarni et al. (2015).

In the latter method, separate word embedding models are trained for each of the time intervals. Since embedding algorithms are inherently stochastic and the resulting embedding sets are invariant under rotation, vectors from these models are not directly comparable and need to be aligned

[1] Code is available at `https://gitlab.com/matej.martinc/semantic_shift_detection`.

in a common space (Kutuzov et al., 2018). To solve this problem, Kulkarni et al. (2015) first suggested a simple linear transformation for projecting embeddings into a common space. Zhang et al. (2016) improved this approach by proposing the use of an additional set of nearest neighbour words from different models that could be used as anchors for alignment. Another approach was devised by Eger and Mehler (2017), who proposed second-order embeddings (i.e., embeddings of word similarities) for model alignment and it was Hamilton et al. (2016a) that showed that these two methods can compliment each other.

Since imperfect aligning can negatively affect semantic shift detection, the newest methods try to avoid it altogether. Rosenfeld and Erk (2018) presented an approach, where the embedding model is trained on word and time representations, treating the same words in different time periods as different tokens. Another solution to avoid alignment is the incremental model fine-tuning, where the model is first trained on the first time period and saved. The weights of this initial model are used for the initialization of the model trained on the next successive time period. The described step of incremental weight initialization is repeated until the models for all time periods are trained. This procedure was first proposed by Kim et al. (2014) and made more efficient by Peng et al. (2017), who suggested to replace the softmax function for the Continuous bag-of-word and Continuous skipgram models with a more efficient hierarchical softmax, and by Kaji and Kobayashi (2017), who proposed an incremental extension for negative sampling.

Recently, a new type of embeddings called contextual embeddings have been introduced. ELMo (Embeddings from Language Models) by Peters et al. (2018) and BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) are the most prominent representatives of this type of contextual embeddings. In this type of embeddings, a different vector is generated for each context a word appears in. These new contextual embeddings solve the problems with word polysemy but have not been used widely in the studies concerning temporal semantic shifts. The only two temporal semantic shift studies we are aware off, that used contextual BERT embeddings, are reported in Hu et al. (2019) and Giulianelli (2019).

In the study by Hu et al. (2019), contextualised BERT embeddings were leveraged to learn a representation for each word sense in a set of polysemic words. Initially, BERT is applied to a diachronic corpus to extract embeddings for tokens that closely match the predefined senses of a specific word. After that, a word sense distribution is computed at each successive time slice. By comparing these distributions, one is able to inspect the evolution of senses for each target word.

In the study by Giulianelli (2019), word meaning is considered as "inherently under determined and contingently modulated in situated language use", meaning that each appearance of a word represents a different word usage. The main idea of the study is to determine how word usages vary through time. First, they fine-tune the BERT model on the entire corpus for domain adaptation and after that they perform diachronic fine-tuning, using the incremental training approach proposed by Kim et al. (2014). After that, the word usages for each time period are clustered with the K-means clustering algorithm and the resulting clusters of different word usages are compared in order to determine how much the word usage changes through time.

The second trend in diahronic semantic change research is a slow shift of focus from researching long-term semantic shifts spanning decades or even centuries to short-term shifts spanning years at most (Kutuzov et al., 2018). For example, a somewhat older research by Sagi et al. (2011) studied differences in the use of English spanning centuries by using the Helsinki corpus (Rissanen et al., 1993). The trend of researching long-term shifts continued with Eger and Mehler (2017) and Hamilton et al. (2016b), who both used the Corpus of Historical American (COHA)[2]. In order to test if existing methods could be applied to detect short-term semantic changes in language, newer research focuses more on tracing short-term socio-cultural semantic shift. Kim et al. (2014) analyzed yearly changes of words in the Google Books Ngram corpus and Kulkarni et al. (2015) analyzed Amazon Movie Reviews, where spans were one year long, and Tweets, where change was measured in months. The most recent exploration of meaning shift over short periods of time that we are aware of, was conducted by Del Tredici et al. (2019), who measured changes of word meaning in online Reddit communities by employing the incremental fine-tuning approach proposed by Kim et al. (2014).

## 3. Methodology

In this section, we present the methodology of the proposed approach by explaining how we obtain time period specific word representations, on which corpora the experiments are conducted, and how we evaluate the approach.

### 3.1. Time specific word representations

Given a set of corpora containing documents from different time periods, we develop a method for locating words with different meaning in different time periods and for quantifying these meaning changes. Our methodology is similar to the approach proposed by Rosenfeld and Erk (2018) since we both construct a time period specific word representation that represents a semantic meaning of a word in a distinct time period.

In the first step, we fine-tune a pretrained BERT language model for domain adaptation on each corpus presented in Section 3.2. Note that we do not conduct any diachronic fine-tuning, therefore our fine-tuning approach differs from the approach presented in Giulianelli (2019), where BERT contextual embeddings were also used, and also from other approaches from the related work that employ the incremental fine-tuning approach (Kim et al., 2014; Del Tredici et al., 2019). The reason behind this lies in the contextual nature of embeddings generated by the BERT model, which are by definition dependent on the time-specific context and therefore, in our opinion, do not require diachronic (time-specific) fine-tuning. We use the English BERT-base-uncased model with 12 attention layers and a hidden layer

[2] http://corpus.byu.edu/coha

size of 768 for experiments on the English corpora, and the multilingual BERT-base-cased model for multilingual experiments[3]. Only one model is used for generating the time period specific word representations in the multilingual setting and not two: one for each language in our experiments, English and Slovenian. We opted for this method in order to generate word representations for both languages that do not need to be aligned in a common vector space and are directly comparable. We only conduct light text preprocessing on the LiverpoolFC corpus, where we remove URLs.

In the next step, we generate time specific representations of words. Each corpus is split into predefined time periods and a set of time specific subcorpora is created for each corpus. The documents from each of the time specific subcorpora are split into sequences of byte-pair encoding tokens (Kudo and Richardson, 2018) of a maximum length of 256 tokens and fed into the fine-tuned BERT model. For each of these sequences of length $n$, we create a sequence embedding by summing the last four encoder output layers. The resulting sequence embedding of size $n$ times *embeddings size* represents a concatenation of contextual embeddings for the $n$ tokens in the input sequence. By chopping it into $n$ pieces, we acquire a representation, i.e., a contextual token embedding, for each word usage in the corpus. Note that these representations vary according to the context in which the token appears, meaning that the same word has a different representation in each specific context (sequence). Finally, the resulting embeddings are aggregated on the token level (i.e., for every token in the corpus vocabulary, we create a list of all their contextual embeddings) and averaged, in order to get a time specific representation for each token in each time period.

Last, we quantitatively estimate the semantic shift of each target word in the period between two time specific representations by measuring the cosine distance between two time specific representations of the same token. This differs from the approach proposed by Giulianelli (2019), where clustering was used as an aggregation method and than Jensen-Shannon divergence was measured, a measure of similarity between probability distributions, to quantify changes between word usages in different time periods.

Another thing to note is that for the experiments on the Brexit news corpus (see Section 3.2.), we conduct the same averaging procedure on the entire corpus (not just on the time specific subcorpus) in order to get a general (not just time specific) representation for each token in the corpus. These general representations of words are used to find the 50 most similar words to the word *Brexit* (see Section 4.2. for further details).

Since the byte-pair input encoding scheme (Kudo and Richardson, 2018) employed by the BERT model does not necessarily generate tokens that correspond to words but rather generate tokens that can sometimes correspond to subparts of words, we also propose the following *on the fly* reconstruction mechanism that allows us to get word representations from byte pair tokens. If a word is split into more than one byte pair tokens, we take an embedding for each byte pair token constituting a word and build a word embedding by averaging these byte pair tokens. The resulting average is used as a context specific word representation.

[3]Although recently a variety of novel transformer language models emerged, some of them outperforming BERT (Yang et al., 2019; Sun et al., 2019), BERT was chosen in this research due to the availability of the pretrained multilingual model which among other languages also supports Slovenian.

## 3.2. Corpora

We used three corpora in our experiments, all of them covering short time periods of eight years or less. The statistics about the datasets are presented in Table 1.

### 3.2.1. LiverpoolFC

The LiverpoolFC corpus is used to compare our approach to a recent state-of-the-art approach proposed by Del Tredici et al. (2019). It contains 8 years of Reddit posts, more specifically the LiverpoolFC subreddit for fans of the English football team. It was created for the task of short-term meaning shift analysis in online communities. The language use in the corpus is specific to a somewhat closed community, which means linguistic innovations are common and non-standard word interpretations are constantly evolving. This makes this corpus very appropriate for testing the models for abrupt semantic shift detection.

We adopt the same procedure as the original authors and split the corpus into two time spans, the first one covering texts ranging from 2011 until 2013 and the second one containing texts from 2017.

### 3.2.2. Brexit news

We compiled the Brexit news corpus to test the ability of our model to detect relative semantic changes (i.e., how does a specific word, in this case *Brexit*, semantically correlate to other words in different time periods) and to test the method on consecutive yearly periods. The subject of Brexit was chosen due to its extensive news coverage over a longer period of time, which allows us to detect possible correlations between the actual events that occurred in relation to this topic and semantic changes detected by the model. The corpus contains about 36.6 million tokens and consists of news articles (more specifically, their titles and content) about Brexit[4] from the RSS feeds of the following news media outlets: Daily Mail, BBC, Mirror, Telegraph, Independent, Guardian, Express, Metro, Times, Standard and Daily Star and the Sun. The corpus is divided into 5 time periods, the first one covering articles about the Brexit before the referendum that occurred on June 23, 2016. The articles published after the referendum are split into 4 yearly periods. The yearly splits are made on June 24 each year and the most recent time period contains only articles from June 24, 2019 until August 23, 2019. The corpus is unbalanced, with time periods of 2016 and 2018 containing much more articles than other splits due to more intensive news reporting. See Table 1 for details.

### 3.2.3. Immigration news

The Immigration news corpus was compiled to test the ability of the model to detect relative semantic changes in a multilingual setting, something that has to our knowledge

[4]Only articles that contain word *Brexit* in the title were used in the corpus creation.

never been tried before. The main idea is to detect similarities and differences in semantic changes related to immigration in two distinct countries with different attitudes and historical experiences about this subject.

The topic of immigration was chosen due to relevance of this topic for media outlets in both countries that were covered, England and Slovenia. The corpus consists of 6,247 English articles and 10,089 Slovenian news articles (more specifically, their titles and content) about immigration[5], is balanced in terms of number of tokens for each language and altogether contains about 12 million tokens. The English and Slovenian documents are combined and shuffled[6] and after that the corpus is divided into 5 yearly periods (split on December 31). The English news articles were gathered from the RSS feeds of the same news media outlets as the news about Brexit, while the Slovenian news articles were gathered from the RSS feeds of the following Slovenian news media outlets: Slovenske novice, 24ur, Dnevnik, Zurnal24, Vecer, Finance and Delo.

| Corpus | Time period | Num. tokens (in millions) |
|---|---|---|
| LiverpoolFC | 2013 | 8.5 |
| LiverpoolFC | 2017 | 11.9 |
| LiverpoolFC | Entire corpus | 20.4 |
| Brexit news | 2011 - 23.6.2016 | 2.6 |
| Brexit news | 24.6.2016 - 23.6.2017 | 10.3 |
| Brexit news | 24.6.2017 - 23.6.2018 | 6.2 |
| Brexit news | 24.6.2018 - 23.6.2019 | 12.7 |
| Brexit news | 24.6.2019 - 23.8.2019 | 2.4 |
| Brexit news | Entire corpus | 36.6 |
| Immigration news | 2015 | 2.2 |
| Immigration news | 2016 | 2.6 |
| Immigration news | 2017 | 2.6 |
| Immigration news | 2018 | 2.6 |
| Immigration news | 2019 | 1.9 |
| Immigration news | Entire corpus | 11.9 |

Table 1: Corpora statistics.

## 3.3. Evaluation

We evaluate the performance of the proposed approach for semantic shift detection by conducting quantitative and qualitative evaluation.

### 3.3.1. Quantitative evaluation

In order to get a quantitative assessment of the performance of the proposed approach, we leverage a publicly available evaluation set for semantic shift detection on the LiverpoolFC corpus (Del Tredici et al., 2019). The evaluation set contains 97 words from the corpus manually annotated with semantic shift labels by the members of the LiverpoolFC subreddit. 26 community members with domain knowledge but no linguistic background were asked to make a binary decision whether the meaning of the word changed between the two time spans (marked as 1) or not (marked as 0) for each of the words in the evaluation set. Each word received on average 8.8 judgements and the average of these judgements is used as a gold standard semantic shift index. Positive examples of meaning shift in this evaluation set can be grouped into three classes according to the type of meaning shift. First are metonymic shifts, which are figures of speech, in which a thing or concept is referred to by the name of something associated with it (e.g., the word *F5* that is initially used as a shortcut for refreshing a page and starts to denote any act of refreshing). Second are metaphorical shifts where the original meaning of a word is widened through analogy (e.g., the word *pharaoh* which is the nickname of an Egyptian football player). Lastly, memes are semantic shifts that occur when a word first used in a humorous or sarcastic way prompts a notable change in word's usage on a community scale (e.g., the first part of the player's surname *Van Dijk* is being used in jokes related to shoes' brand *Vans*).

We measure Pearson correlation between the semantic shift index and the model's semantic shift assessment for each of the words in the evaluation set in order to be able to directly compare our approach to the one presented in Del Tredici et al. (2019), where the same evaluation procedure was employed. As explained in Section 3.1., we obtain semantic shift assessments by measuring the cosine distance between two time specific representations of the same token.

### 3.3.2. Qualitative evaluation

For the Brexit and Immigration news corpora, manually labeled evaluation sets are not available, therefore we were not able to quantitatively assess the approach's performance on these two corpora. For this reason, the performance of the model on these two corpora is evaluated indirectly, by measuring how does a specific word of interest semantically correlate to other seed words in a specific time period and how does this correlation vary through time. The cosine distance between the time specific representation of a word of interest and the specific seed word is used as a measure of semantic relatedness. We can evaluate the performance of the model in a qualitative way by exploring if detected differences in semantic relatedness (i.e., relative semantic shifts) are in line with the occurrences of relevant events which affected the news reporting about Brexit and Immigration, and also the findings from the academic studies on these topics. This is possible because topics of Brexit and Immigration have been extensively covered in the news and several qualitative analyses on the subject have been conducted.

The hypothesis that justifies this type of evaluation comes from structural linguistics and states that word meaning is a relational concept and that words obtain meaning only in relation to their neighbours (Matthews and Matthews, 2001). According to this hypothesis, the change in the word's meaning is therefore expressed by the change in

[5]The corpus contains English articles that contain words *immigration*, *immigrant* or *immigrants* in the title and Slovenian articles that contain Slovenian translations of these words in either title or content.

[6]Shuffling is performed to avoid the scenario where all English documents would be at the beginning of the corpus and all Slovenian documents at the end, which would negatively affect the language model fine-tuning.

semantic relatedness to other neighbouring words. Neighbouring seed words to which we compare the word of interest for the Brexit news corpus are selected automatically (see Section 4.2. for details) while for the Immigration news corpus, the chosen seed words are concepts representing most common aspects of the discourse about immigration (see Section 4.3. for details).

## 4. Experiments

In this section we present details about conducted experiments and results on the LiverpoolFC, Brexit and Immigration corpora.

### 4.1. LiverpoolFC

In this first experiment, we offer a direct comparison of the proposed method to the state-of-the-art approach proposed by Del Tredici et al. (2019). In their study, they use a Continuous Skipgram model proposed by Mikolov et al. (2013) and employ the incremental model fine-tuning approach first proposed by Kim et al. (2014). In the first step, they create a large Reddit corpus (with about 900 million tokens) containing Reddit post from the year 2013 and use it for training the domain specific word embeddings. The embeddings of this initial model are used for the initialization of the model trained on the next successive time period, LiverpoolFC 2013 posts, and finally, the embeddings of the LiverpoolFC 2013 model are used for the initialization of the model trained on the LiverpoolFC 2017 posts. We, on the other hand, do not conduct any additional domain adaptation on a large Reddit corpus and only fine-tune the BERT model on the LiverpoolFC corpus, as already explained in Section 3..

First, we report on the results of the diachronic semantic shift detection for 97 words from the LiverpoolFC corpus that were manually annotated with semantic shift labels by members of the LiverpoolFC subreddit (see Section 3.2.1. for more details on the annotation and evaluation procedures). Overall, our proposed approach yields almost identical positive correlation between cosine distance between 2013 and 2017 word representations and semantic shift index as in the research conducted by Del Tredici et al. (2019). We observe the Pearson correlation of 0.47 ($p < 0.001$) while the original study reports Pearson correlation of 0.49.

On the other hand, there are also some important differences between the two methods. Our approach (see Figure 1) proves to be more conservative when it comes to measuring the semantic shift in terms of cosine distance. In the original approach, the cosine distance of up to 0.6 is measured for some of the words in the corpus, while we only observe the differences in cosine distance of up to 0.3 (for the word *roast*). This conservatism of the model results in less false positive examples (i.e., detected semantic shifts that were not observed by human annotators) compared to the original study, but also results in more false negative examples (i.e., unrecognised semantic shifts that were recognized by human annotators)[7]. An example of a false negative detection by the system proposed by Del Tredici et al. (2019) is the word *lean*. An example of a false positive detection by the system proposed by Del Tredici et al. (2019) that was correctly identified by our system as a word with unchanged semantic context is the word *stubborn*. On the other hand, our system also manages to correctly identify some of the words that changed the most that were misclassified by the system proposed by Del Tredici et al. (2019). An example of this is the word *Pharaoh*.

There are also some similarities between the two systems. For example, the word *highlighter* is correctly identified as a word that changed meaning by both systems. With the exception of *Pharaoh*, we also notice similar tendencies of both systems to misclassify as false negatives words that fit into the category of so-called metaphorical shifts (i.e., widening of the original meaning of a word through analogy). Examples of these words would be *snake*, *thunder* and *shovel*. One explanation for this misclassification that was offered by Del Tredici et al. (2019) is the fact that many times the metaphoric usage is very similar to the literal one, therefore preventing the model to notice the difference in meaning[8].

### 4.2. Brexit news

Here we asses the performance of the proposed approach for detecting sequential semantic shift of words in short-term yearly periods. More specifically, we explore how time specific seed word representations in different time periods change their semantic relatedness to the time specific word representation of the word *Brexit*. The following procedure is conducted. First, we find 50 words most semantically related to the general non-time specific representation of *Brexit* according to their non-time specific general representations. Since the initial experiments showed that many of the 50 most similar words are in fact derivatives of the word *Brexit* (e.g., *brexitday*, *brexiters*...) and therefore not that relevant for the purpose of this study (as their meaning is fully dependant on the concept from which they derived), we first conduct an additional filtering according to the normalized Levenshtein distance defined as:

$$normLD = 1 - (LD/max(len(w1), len(w2))),$$

where $normLD$ stands for normalized Levenshtein difference, $LD$ for Levenshtein difference, $w1$ is *Brexit* and $w2$ are other words in the corpus. Words for which normalized Levenshtein difference is bigger than 0.5 are discarded and out of the remaining words we extract 50 words most semantically related to *Brexit* according to the cosine similarity. [9]

Out of these 50 words, we find ten words that changed the most in relation to the time specific representation of the

[7] Expressions *false positive* and *false negative* are used here to improve the readability of the paper and should not be interpreted in a narrow context of binary classification.

[8] For example, *shovel* is used in a context where the team is seen as a train running through the season, and the fan's job is to contribute in a figurative way by shoving the coal into the train boiler. Therefore, the word *shovel* is used in sentences like *You boys know how to shovel coal.*

[9] The normalized Levenshtein difference treshold of 0.5 and the number of most semantically similar words were chosen empirically.

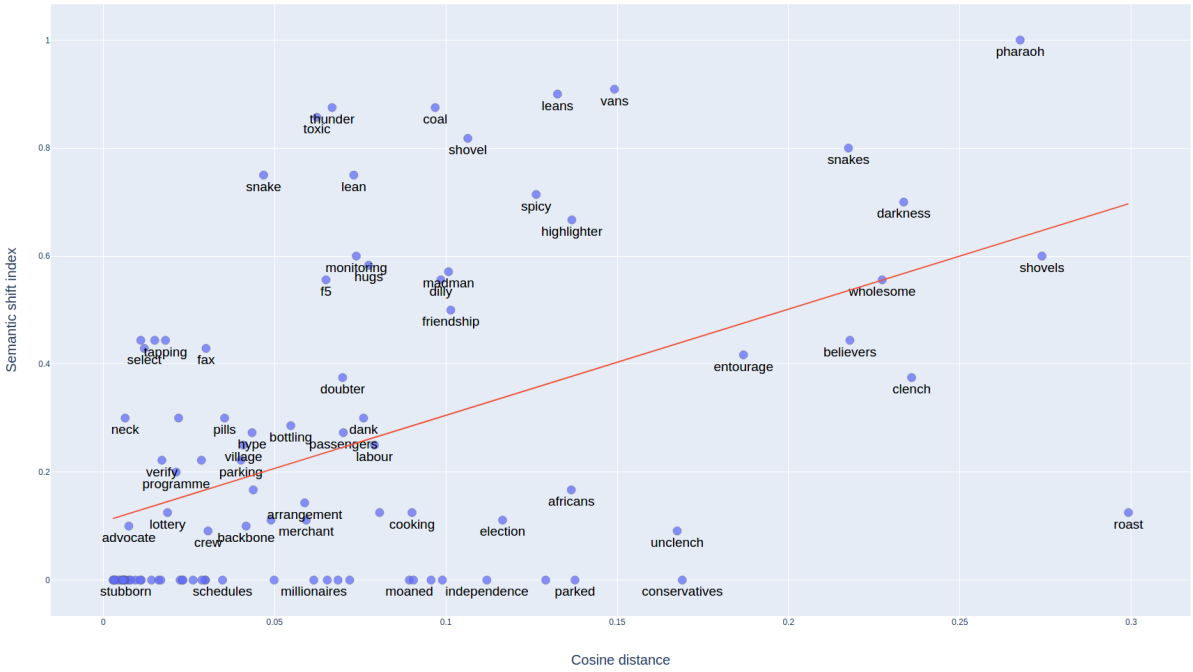



Figure 1: Semantic shift index vs. cosine distance in the LiverpoolF1 evaluation dataset.

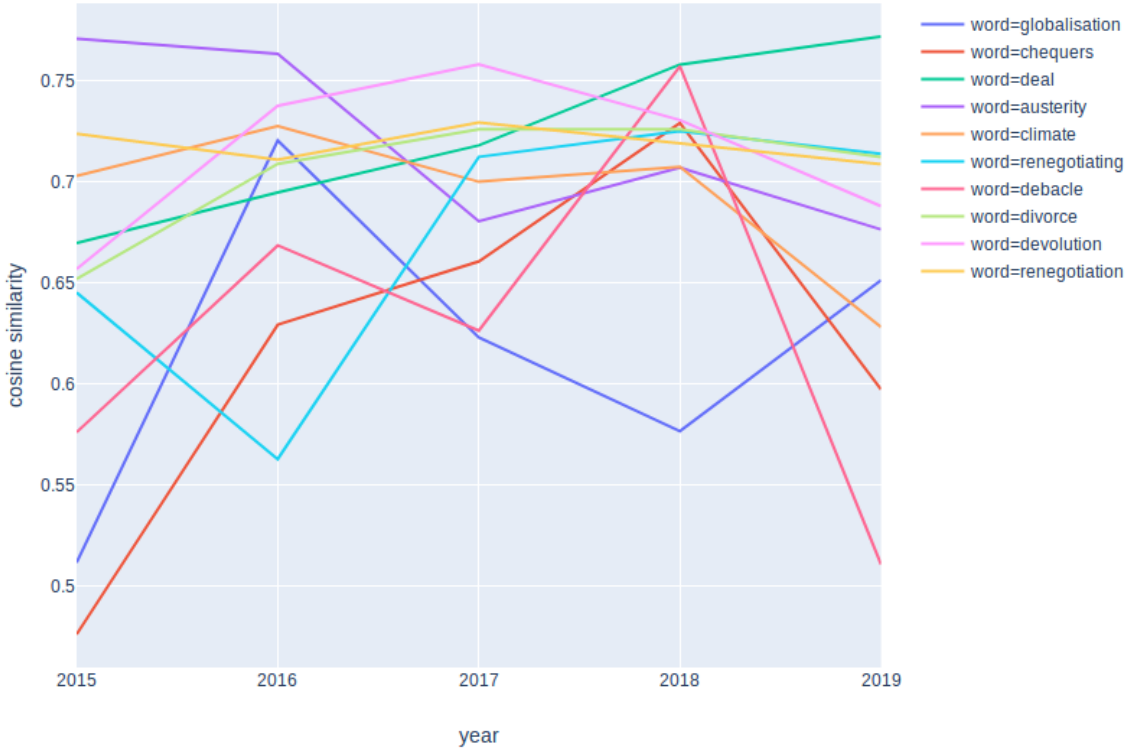



Figure 2: The relative diachronic semantic shift of the word *Brexit* in relation to ten words that changed the most out of 50 closest words to *Brexit* according to the cosine similarity.

word *Brexit* with the following equation:

$$MC = abs(CS(w1_{2015}, w2_{2015}) - CS(w1_{2019}, w2_{2019}))$$

where *MC* stands for meaning change, *CS* stands for cosine similarity, $w1_{year}$ is a year specific representation of the word *Brexit* and $w2_{year}$ are year specific representations of words related to *Brexit*.

The resulting 10 seed words are used to determine the relative diachronic semantic shift of the word *Brexit* as explained in Section 3.3. Figure 2 shows the results of the experiments. We can see that the word *deal* is constantly becoming more and more related to *Brexit*, from having a cosine similarity to the word *Brexit* of about 0.67 in 2015 to having a cosine similarity of about 0.77 in 2019. This is consistent with our expectations. The biggest overall difference of about 0.14 in semantic relatedness can be observed for the word *globalisation*, having not been very related to *Brexit* before the referendum in year 2016 (with the cosine similarity of about 0.52) and than becoming very related to the word *Brexit* in a year after the referendum (with cosine similarity of about 0.72). After that, we can observe another drop in similarity in the following two years and then once again a rise in similarity in 2019. This movement could be at least partially explained by the post-referendum debate on whether UK's *Leave* vote could be seen as a vote against globalisation (Coyle, 2016).

A sudden rise in semantic relatedness between words *Brexit* and *devolution* in years 2016 and 2017 could be explained by a still quite relevant question of how UK's withdrawal from the EU will affect its structures of power and administration (Hazell and Renwick, 2016). We can also observe a quite sudden drop in semantic relatedness between the words *Brexit* and *austerity* in year 2017, one year after the referendum. It is possible, that the debate on whether UK's *leave* vote was caused by austerity-induced welfare reforms proposed by the UK government in 2010 (Fetzer, 2019) has been calming down. Another interesting thing to note is the enormous drop of about 0.25 in cosine similarity for the word *debacle* after June 23 2019, which has gained the most in terms of semantic relatedness to the word *Brexit* in 2018. It is possible that this gain is related to the constant delays in the UK's attempts of leaving the EU).

Some findings of the model are harder to explain. For example, according to the model, the talk about the *renegotiating* in the context of *Brexit* has not been very common in years 2015 and 2016 and than we can see a major rise of about 0.15 in cosine similarity in year 2017. On the other hand, an almost identical word *renegotiation* has kept a very steady cosine similarity of about 0.72 throughout an entire five year period. We also do not have an explanation for a large drop in semantic relatedness in 2019 between words *chequers* and *Brexit*, and *climate* and *Brexit*.

### 4.3. Immigration news

Here we asses the performance of the proposed approach in a multilingual English-Slovenian setting. Since the main point of these experiments is to detect differences and similarities in relative semantic shift in two distinct languages, we first define English-Slovenian word pairs that arguably represent some of the most common aspects of the discourse about immigration (Martinez Jr and Lee, 2000; Borjas, 1995; Heckmann and Schnapper, 2016; Cornelius and Rosenblum, 2005). These English-Slovenian matching translations are *crime-kriminal*, *economy-gospodarstvo*, *integration-integracija* and *politics-politika*. We measure the cosine similarity between time specific vector representations of each word in the word pair and a time specific vector representation of a word *immigration*.

The results of the experiments are presented in Figure 3. First thing one can note is that in most cases and in most years English and Slovenian parts of a word pair have a very similar semantic correlation to a word *immigration*, which suggest that the discourse about immigration is quite similar in both countries. The similarity is most apparent for the word pair *crime-kriminal* and to a slightly lesser extent for the word pair *politics-politika*. On the other hand, not much similarity in relation to a word *immigration* can be observed for an English and Slovenian words for economy. This could be partially explained with the fact that Slovenia is usually not a final destination for modern day immigrants (who therefore do not have any economical impact on the country) and serves more as a transitional country (Garb, 2018), therefore immigration is less likely to be discussed from the economical perspective.

Figure 3 also shows some interesting language specific yearly meaning shifts. The first one is the rise in semantic relatedness between the word *immigration* and the English word *politics* in 2016. This could perhaps be related to the Brexit referendum which occurred in the middle of the year 2016 and the topic of *immigration* was discussed by politicians from both sides of the political spectrum extensively in the referendum campaign.

Another interesting yet currently unexplainable yearly shift concerns Slovenian and English words for *integration* in 2019. While there is a distinct fall in semantic relatedness between words *integration* and *immigration*, we can on the other hand observe a distinct rise in semantic relatedness between words *integracija* and *immigration*.

## 5. Conclusion

We presented a research on how contextual embeddings can be leveraged for the task of diachronic semantic shift detection. A new method that uses BERT embeddings for creating time specific word representations was proposed and we showcase the performance of the new approach on three distinct corpora, LiverpoolFC, Brexit news and Immigration news.

The proposed method shows comparable performance to the state-of-the-art on the LiverpoolFC corpus, even though domain adaptation was performed only on the corpus itself and no additional resources (as was the case in the study by Del Tredici et al. (2019)) were required. This shows that the semantic knowledge that BERT model acquired during its pretraining phase can be successfully transferred into domain specific corpora. This is welcome from the stand point of reduced time complexity (since training BERT or most other embedding models from scratch is very time consuming) and it also makes our proposed method appropriate for

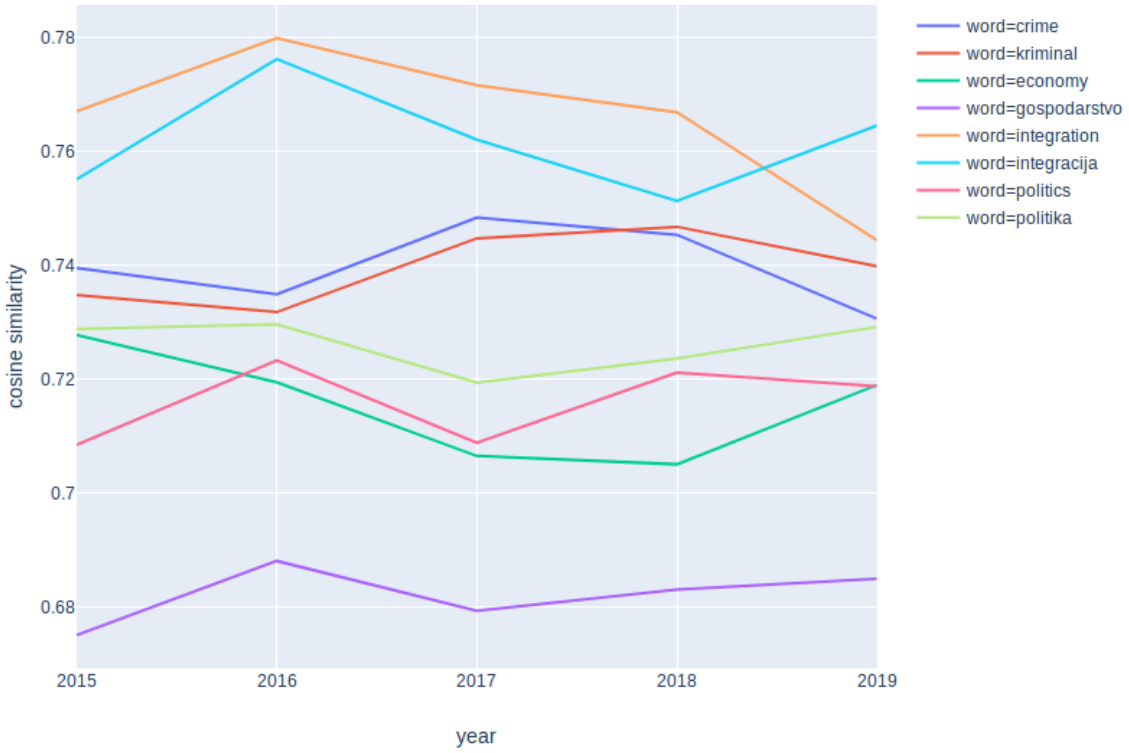



Figure 3: The relative diachronic semantic shift of the word *immigration* in relation to English-Slovenian word pairs crime-kriminal, economy-gospodarstvo, integration-integracija and politics-politika.

detecting meaning shifts in domains for which large corpora are not available.

Experiments on the Brexit news corpus are also encouraging, since detected relative semantic shifts are somewhat in line with the occurrence of different events which affected the news reporting about Brexit in different time periods. Same could be said for the multi-lingual experiments conducted on the English-Slovenian Immigration news corpus, which is to our knowledge the first attempt to compare parallel meaning shifts in two different languages, and opens new paths for multilingual news analysis.

On the other hand, a lot of further work still needs to be done. While the results on the Brexit and Immigration news corpora are encouraging, a more thorough evaluation of the approach would be needed. This could either be done in comparison to a qualitative discourse analysis or by a quantitative manual evaluation, in which changes detected by the proposed method would be compared to changes identified by human experts with domain knowledge, similar as in Del Tredici et al. (2019).

The method itself could also be refined or improved in some aspects. While we demonstrated that averaging embeddings of word occurrences in order to get time specific word representations works, we did not experiment with other grouping techniques, such as taking median word representation instead of an average or by using weighted averages. Another option would also be to further develop clustering aggregation techniques, similar as in Giulianelli (2019). While these methods are far more computationally demanding and less scalable than averaging, they do have an advantage of better interpretability, since clustering of word usages into a set of distinct clusters resembles the manual approach of choosing the word's contextual meaning from a set of predefined meanings.

## 6. Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The authors acknowledge also the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103) and the project TermFrame - Terminology and Knowledge Frames across Languages (No. J6-9372). We would also like to thank Lucija Luetic for useful discussions.

## 7. Bibliographical References

Bloomfield, L. (1933). *Language history: from Language (1933 ed.).* Holt, Rinehart and Winston.

Borjas, G. J. (1995). The economic benefits from immigration. *Journal of economic perspectives*, 9(2):3–22.

Cornelius, W. A. and Rosenblum, M. R. (2005). Immigration and politics. *Annu. Rev. Polit. Sci.*, 8:99–119.

Coyle, D. (2016). Brexit and globalisation. *Brexit Beckons: Thinking ahead by leading economists*, page 23.

Del Tredici, M., Fernández, R., and Boleda, G. (2019). Short-term meaning shift: a distributional exploration. 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 2069–2075.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eger, S. and Mehler, A. (2017). On the linearity of semantic change: Investigating meaning variation via dynamic graph models. *arXiv preprint arXiv:1704.02497*.

Fetzer, T. (2019). Did austerity cause brexit? *American Economic Review*, 109(11):3849–86.

Garb, M. (2018). Coping with the refugee and migrant crisis in slovenia: the role of the military. *Defense & Security Analysis*, 34(1):3–15.

Giulianelli, M. (2019). *Lexical Semantic Change Analysis with Contextualised Word Representations*. University of Amsterdam - Institute for logic, Language and computation.

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 Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Hazell, R. and Renwick, A. (2016). Brexit: its consequences for devolution and the union. *UCL Constitution Unit Briefing Paper, The Constitution Unit*, 19:4.

Heckmann, F. and Schnapper, D. (2016). *The integration of immigrants in European societies: National differences and trends of convergence*, volume 7. Walter de Gruyter GmbH & Co KG.

Hilpert, M. and Gries, S. T. (2008). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

Hu, R., Li, S., and Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, 37(1):77–96.

Kaji, N. and Kobayashi, H. (2017). Incremental skip-gram model with negative sampling. *arXiv preprint arXiv:1704.03956*.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

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.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Martinez Jr, R. and Lee, M. T. (2000). On immigration and crime. *Criminal justice*, 1(1):486–524.

Matthews, P. H. and Matthews, P. (2001). *A short history of structural linguistics*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Peng, H., Li, J., Song, Y., and Liu, Y. (2017). Incrementally learning the hierarchical softmax function for neural language models. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Rissanen, M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., Nevalainen, T., and Raumolin-Brunberg, H. (1993). The helsinki corpus of english texts. *Kyttö et. al*, pages 73–81.

Rosenfeld, A. and Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484.

Sagi, E., Kaufmann, S., and Clark, B. (2011). Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, 73:161–183.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Zhang, Y., Jatowt, A., Bhowmick, S. S., and Tanaka, K. (2016). The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807.

# Appendix B: Capturing evolution of word usage: Just add more clusters

## Capturing Evolution in Word Usage: Just Add More Clusters?

Matej Martinc*
matej.martinc@ijs.si
Jozef Stefan Institute
Slovenia

Syrielle Montariol*
syrielle.montariol@limsi.fr
LIMSI - CNRS, Univ. Paris-Sud, Univ. Paris-Saclay, Societé Générale
France

Elaine Zosa*
elaine.zosa@helsinki.fi
University of Helsinki
Finland

Lidia Pivovarova
lidia.pivovarova@helsinki.fi
University of Helsinki
Finland

### ABSTRACT

The way the words are used evolves through time, mirroring cultural or technological evolution of society. Semantic change detection is the task of detecting and analysing word evolution in textual data, even in short periods of time. In this paper we focus on a new set of methods relying on contextualised embeddings, a type of semantic modelling that revolutionised the NLP field recently. We leverage the ability of the transformer-based BERT model to generate contextualised embeddings capable of detecting semantic change of words across time. Several approaches are compared in a common setting in order to establish strengths and weaknesses for each of them. We also propose several ideas for improvements, managing to drastically improve the performance of existing approaches.

### CCS CONCEPTS

• **Computing methodologies** → **Lexical semantics**; *Cluster analysis*; • **Information systems** → *Language models.*

### KEYWORDS

Semantic Change, Contextualised Embeddings, Clustering

**ACM Reference Format:**
Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing Evolution in Word Usage: Just Add More Clusters?. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366424.3382186

### 1 INTRODUCTION

The large majority of data on the Web is unstructured. Amongst it, textual data is an invaluable asset for data analysts. With the large increase in volume of interaction and overall usage of the Web, more and more content is digitised and made available online, leading to a huge amount of textual data from many time periods becoming accessible. However, textual data are not necessarily homogeneous as they rely on a crucial element that evolves throughout time: language. Indeed, a language can be considered as a dynamic system where word usages evolve over time, mirroring cultural or technological evolution of society [1].

In linguistics, *diachrony* refers to the study of temporal variations in the use and meaning of a word. While analysing textual data from the Web, detecting and understanding these changes can be done for two primary goals. First, it can be used directly for linguistic research or social analysis, by interpreting the reason of the semantic change and linking it to real-world events, and by analysing trends, topics and opinions evolution [9]. Second, it can be used as a support for many tasks in Natural Language Processing (NLP), from text classification to information retrieval conducted on a temporal corpora where semantic change might occur.

To tackle semantic change, models usually rely on word embeddings, which summarise all senses and usages of a word within a certain time period into one vector. Measuring the distance between these vectors across time periods is used to detect and quantify the differences in meaning. But these methods do not take into consideration that most words have multiple senses, since all word usages are aggregated into a single static word embedding. Contextualised embedding models such as BERT [5] are capable of generating a separate vector representation for each specific word usage, making them more suitable for this task.

The goal of this paper is to establish the best way to detect semantic change in a temporal corpus by capitalising on BERT contextualised embeddings. First, several approaches for semantic shift detection from the literature are compared in a common setting in order to establish strengths and weaknesses of each specific method. Second, several improvements are presented, which manage to drastically improve the performance of existing approaches. Our code and models are publicly available[1].

### 2 RELATED WORK

A large majority of methods for semantic shift detection leverage dense word representations, i.e. embeddings. Word-frequency methods for detecting semantic shift that were popular in earlier studies [13, 16], are now rarely used. The detailed overview of the field could be found in recent surveys [22, 27, 28].

*The authors contributed equally to this research.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
*WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan*
© 2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3382186

[1] https://github.com/smontariol/AddMoreClusters

### 2.1 Static Word Embeddings for Semantic Change

The first research that employed word embeddings for semantic shift detection was conducted by [18]. The main idea was to train a separate embedding model for each time period. Since embedding algorithms are inherently stochastic and the resulting embedding sets are invariant under rotation, a procedure that makes these models comparable is needed. To solve this problem, they proposed the incremental model fine-tuning approach, where the weights of the model, trained on a certain time period, are used to initialize weights of a model trained on the next successive time period. Some improvements of the approach were later proposed by [24], who replaced the softmax function for the continuous skipgram model with a more efficient hierarchical softmax, and by [17], who proposed an incremental extension for negative sampling.

An alternative approach was proposed in [19], where embedding models trained on different time periods were aligned in a common vector space after the initial training using a linear transformation for the alignment. The approach was upgraded [31] by using a set of nearest neighbour words as anchors for the alignment.

The third alternative for semantic shift detection with static word embeddings is to treat the same words in different time periods as different tokens in order to get time specific word representations for each time period [6, 26]. Here, only one embedding model needs to be trained and no aligning is needed.

### 2.2 The Emergence of Contextualised Embeddings

While in static word embedding models each word from the predefined vocabulary is presented as a unique vector, in contextualised embeddings a separate vector is generated for each word mention, i.e. for each context the word appears in. The two most widely used contextual embeddings models are ELMo (Embeddings from LanguageModels [25]) and a more recent BERT (Bidirectional Encoder Representations from Transformers [5]). The approach of using contextual embeddings for semantic shift detection is fairly novel; we are aware of three recent studies that employed it.

In the first study, contextualised embeddings were applied in a controlled way [15]: for a set of polysemic words, a representation for each sense is learned using BERT. Then pretrained BERT is applied to a diachronic corpus, extracting token embeddings, that are matched to the closest sense embedding. Finally, the proportions for each sense are computed at each successive time slice, revealing the evolution of the distribution of senses for each target word. This method requires that the set of senses of each target word is known beforehand.

Another possibility is clustering all contextual embeddings for a target word into clusters representing the word senses or usages in a specific time periods [10]. K-means clustering and BERT contextual embeddings were used in this study. In addition, the incremental training approach proposed by [18] was used for diachronic fine-tuning of the model. Jensen-Shannon divergence (JSD), a measure of similarity between probability distributions, was used to quantify changes between word usages in different time periods. They also tested if domain adaptation of the model would improve the results of their approach by fine-tuning the model on an entire corpus rather than on specific time periods, however this yielded no performance improvements.

In the third, even more recent study, contextual embeddings for a specific word in a specific time period were averaged in order to generate a time specific word representation for each word in each period [23]. BERT embeddings are used in the study and cosine distance is used for measuring the difference between word representations in different time periods.

## 3 DATA

We rely on a small human-annotated dataset [12] to conduct the evaluation. The dataset consists of 100 words from various frequency ranges, labelled by five annotators according to the level of semantic change between the 1960s and the 1990s. They use a 4-points scale from "0: no change" to "3: significant change", and the inter-rater agreement was 0.51 ($p < 0.01$, average of pair-wise Pearson correlations). The most significantly changed words from the dataset are, for example, *user* and *domain*; words for which the meaning remain intact, are for example *justice* and *chemistry*. This dataset is a valuable resource and has been used to evaluate methods for measuring semantic change in previous research [7, 10]. Following previous work, we use the average of the human annotations as semantic change score. For evaluation, we compute Pearson and Spearman rank correlations between this score and a model output. The notion of the best model is based on Spearman correlations.

To train the models we use the Corpus of Historical American English (COHA) [2]. It contains more than 400 million words of text from the 1810s-2000s. As a historical corpus, it is smaller than the widely used Google books corpus [3] but it has the advantage that data from each decade are balanced by genre—fiction, magazines, newspapers, and non-fiction texts, gathered from various Web sources. We focus our experiments on the most recent data in this corpus, from the 1960s to the 1990s (1960s has around 2.8 million and 1990s 3.3 million words), to match the manually annotated data. The fine-tuning of the model is also done only on this subset.

## 4 METHODOLOGY

### 4.1 Context-dependent Embeddings

BERT is a neural model based on the transformer architecture [29]. It relies on a transfer learning approach proposed by [14], where in the first step the network is pretrained as a language model on large corpora in order to learn general contextual word representations. This is usually followed by a task specific fine-tuning step e.g., classification or, in our case, domain adaptation. BERT's novelty is an introduction of a new pretraining learning objective, a *masked language model*, where a percentage of words from the input sequence is masked in advance, and the objective is to predict these masked words from an unmasked context. This allows BERT to leverage both left and right context, meaning that a word $w_t$ in a sequence is not determined just from its left sequence $w_{1:t-1} = [w_1, ..., w_{t-1}]$—as is the case in the traditional language modelling task—but also from its right word sequence $w_{t+1:n} = [w_{t+1}, ..., w_{t+n}]$.

[2] https://www.english-corpora.org/coha/
[3] http://googlebooks.byu.edu/

In our experiments we use the English BERT-base-uncased model with 12 attention layers and a hidden layer of size 768, which was pretrained on the Google Books Corpus [11] (800M words) and Wikipedia (2,500M words). For some of the experiments (see Table 1), we further fine-tune this model (as a *masked language model*) for up to 10 epochs on the COHA subcorpus described in Section 3 for domain adaptation.

Note that our fine-tuning approach deviates from the approaches presented in some of the related work [10] and we do not conduct any diachronic fine-tuning of the model using the incremental training approach similar to [18]. The hypothesis is that this step is not necessary due to contextual nature of embeddings generated by the model, which by definition are dependent on the context that is always time-specific.

Since we are using a pre-trained model we have to apply the BERT tokenization, which is based on byte-pair encodings [30]. In order to acquire contextual embeddings, the corpus documents are first split into sentences; each sentence is limited to 512 tokens and fed into the BERT model. A sequence embedding is generated for each of these sequences by summing last four encoder output layers of BERT[4]. Finally, this sequence embedding of size $sequence\ length \times embeddings\ size$ is cut into pieces, to get a separate contextual embedding for each token in the sequence.

### 4.2 Target Words Selection

In any practical application of semantic change detection, performing clustering for every word in the corpus would not be feasible in terms of computing time. Thus, we investigate several scalable metrics as a preliminary step to identify a set of words that may have undergone semantic change.

A first set of metrics relies on the computation of a *variation* measure, similarly to [20]. Variation is the cosine distance between each token embedding and a *centroid*, i.e. an average token embedding for a given word. The mean of these cosine distances is the *variation coefficient* of a word. The intuition is that for words that have many different senses and usages, the distance to the centroid would be higher than for words that are monosemous. However, this method does not make distinction between words that gain (loose) sense and polysemous words that stay stable across time.

To measure an evolution of word variation, we compute the variation coefficient inside each time slice $t$. Then, we take the average difference from one time step to another. This measure aims at detecting words that undergo changes in their level of polysemy. For example, in a corpus divided into T time slices:

$$Variation\ by\ time\ slice = \frac{\sum_{t=t_0}^{T} |Variation_t - Variation_{t-1}|}{T},$$

The second set of metrics relies on *averaging* all token embeddings at each time slice, and using the cosine distance as a measure of semantic drift between time slices. The total drift is the cosine distance between the average of token representations of the first time slice and of the last time slice. It represents the amount of change a word has undergone from the first to the last period, without taking into account the variations in between. The *averaging by time slice* computes the mean of the drifts from each time step to the next one, in order to measure the successive changes of word usage.

To evaluate and compare these measures we use all hundred words from the test set. In practice it is possible to choose a threshold (as a fraction of the size of the full vocabulary) to get a list of target words. Then, the heavier clustering techniques can be applied to this list.

### 4.3 Embeddings Clustering

The goal of the clustering step is to group the word occurrences by similar vector representation. Then JSD is used to compare cluster distribution across time periods, same as in [10]. The intuition is the following: if, for instance, a word acquired a novel sense in the latter time period, then a cluster corresponding to this sense only consists of word usages from this period but not the earlier ones, which would be reflected by a higher divergence. However, a cluster does not necessarily correspond to a precise sense of the word. Each cluster would rather represent a specific usage or context. Moreover, a word may completely change its context without changing the meaning. Consequently, determining the number of clusters is a tricky part.

For clustering we used k-means with various values for $k$ and affinity propagation [8]. Affinity propagation has been previously used for various linguistic tasks, such as word sense induction [2, 21]. Affinity propagation is based on incremental graph-based algorithm, partially similar to PageRank. Its main strength is that number of clusters is not defined in advance but inferred during training. We also experiment with the approach inspired by [3], where clusters with less than two members are considered weak and merged with the closest strong cluster, i.e. clusters with more than two members.[5] We refer to this method as two-stage clustering.

## 5 EXPERIMENTS

We focus our analysis on comparing the various clustering approaches and the metrics to detect semantic change. Table 1 shows the Pearson and Spearman correlations between the models' outputs and the human-annotated drifts. We also report Silhouette scores for clustering.

We use a pretrained version of BERT [6] and BERT fine-tuned on the COHA subcorpus for up to 10 epochs. We make use of the Scikit-learn implementation of k-means and affinity propagation [7]. For k-means, we set the number of clusters $k$ and use default parameters for the rest. Similarly, for affinity propagation, we use the default parameters set by the library.

A specificity of BERT is the representation of words with byte-pair encodings [30]. Thus, some words can be divided into several sub-parts; for example, in our list of hundred target words for evaluation, *sulphate* is divided into two byte-pairs *sul* and *##phate*, where *##* denotes the splitting of the word. This is also true for the words *medieval*, *extracellular* and *assay*. We decided to exclude these words from our analysis. Thus, strictly speaking our results

[4] We refer the reader to the original implementation of transformer in [29] for a detailed overview of each component in the architecture.

[5] Note that procedure in [3] is more complex: they first find one or more number of representatives for each datapoint and then clustering is applied over representatives, while in our work clustering is done over the instances themselves.

[6] https://pytorch.org/hub/huggingface_pytorch-transformers/

[7] https://scikit-learn.org/stable/modules/clustering.html

**Table 1: Correlations between detected semantic change and manually annotated list of semantic drifts [12] between 1960s and 1990s.**

| Method | Pearson | Spearman | Silhouette |
|---|---|---|---|
| **Related work** | | | |
| **Gulardova & Baroni, 2011** [12] | 0.386 | - | - |
| **Frermann & Lapata, 2016** [7] | - | 0.377 | - |
| **Giulianelli, 2019** [10] | 0.231 | 0.293 | - |
| **Kutuzov, 2020** [20] | 0.233 | 0.285 | - |
| **Pretrained BERT** | | | |
| *Target word selection* | | | |
| **Variation** | 0.070 | 0.015 | - |
| **Variation by decade** | 0.239 | 0.303 | - |
| **Averaging by decade** | 0.295 | 0.272 | - |
| **Averaging** | **0.354** | **0.349** | - |
| *Clustering* | | | |
| **k-means, k = 3** | 0.461 | 0.444 | 0.104 |
| **k-means, k = 5** | 0.476 | 0.443 | 0.096 |
| **k-means, k = 7** | 0.485 | 0.434 | 0.091 |
| **k-means, k = 10** | 0.478 | 0.443 | 0.086 |
| **2-stage clustering, Aff. propagation** | 0.530 | 0.485 | - |
| **Affinity propagation** | **0.548** | **0.486** | 0.039 |
| **Fine-tuned BERT for 5 epochs** | | | |
| *Target word selection* | | | |
| **Averaging** | 0.317 | 0.341 | - |
| *Clustering* | | | |
| **k-means, k=3** | 0.411 | 0.392 | **0.105** |
| **k-means, k=5** | 0.539 | 0.508 | 0.098 |
| **k-means, k=7** | 0.526 | 0.491 | 0.092 |
| **k-means, k=10** | 0.500 | 0.466 | 0.088 |
| **k-means, k=100** | 0.315 | 0.337 | 0.042 |
| **2-stage clustering, Aff. propagation** | 0.554 | 0.502 | - |
| **Affinity propagation** | **0.560** | **0.510** | 0.043 |

are not directly comparable to some of the other approaches in the literature that do not employ BERT.

At the top of Table 1 we overview all previous work on the same test set. To train the models, [13] used GoogleBooks Ngrams, [8] used an extended COHA corpus, and both [11] and [21] used a subcorpus of COHA, identical to the one used in our experiments. In fact, the setting in [11] is quite similar to our work, though our best model performance is much higher than in [11]; we will further discuss this discrepancy in Section 6.

As can be seen in Table 1, among all metrics used for target word selection averaging yields the highest correlation with the human annotations. This intuitively makes sense since averaging measures semantic drift between the first and the last time step and the evaluation dataset was annotated by only considering the first and the last decade. Variation by decade also shows good results; it is a measure of the evolution of the level of variation of a word usage through time.

As can be seen in Table 1 affinity propagation on the fine-tuned BERT model yields the highest Spearman rank correlation. Results obtained using pretrained and fine-tuned models are consistent: in both runs averaging yields lower performance than clustering and affinity propagation is the best clustering method. Two-stage clustering works better than k-means but slightly worse than affinity propagation.

Fine-tuning BERT improves all models except for k-means with 3 clusters and averaging—we do not yet have a clear explanation for that exception.

To conclude, clustering fine-tuned embeddings using affinity propagation yields the best results, with a Pearson correlation with human annotation of 0.56. To evaluate the success of this result, we can use the value of the inter-rater agreement during the annotation process, which was 0.51, computed using the average of pair-wise Pearson correlations [12]. This highlights the difficulty of the task and the performance of the best method.

## 6 DISCUSSION

### 6.1 Error Analysis

We manually checked few examples by choosing the words that have less mentions in the corpus to be able to look through all sentences containing the word. One of the tricky cases for our model is the word *neutron*: according to the manual annotation,

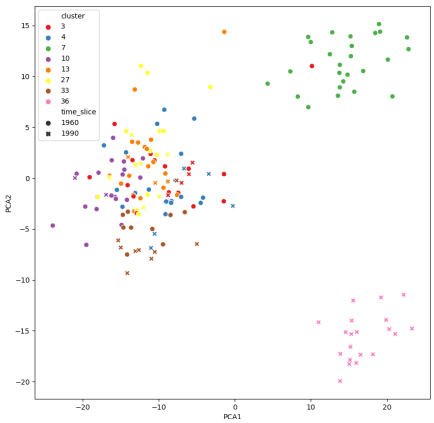



**Figure 1: 2D PCA visualization for the biggest clusters obtained for word *neutron*.**

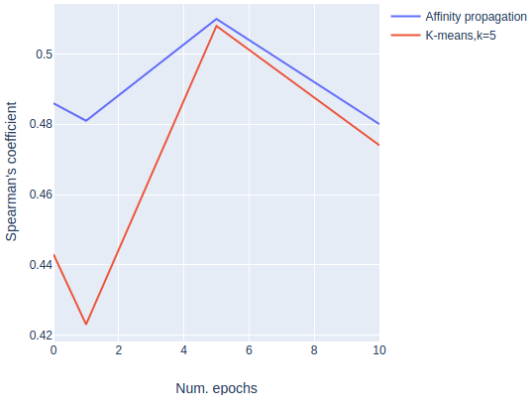



**Figure 2: Impact of BERT fine-tuning on the performance of two distinct aggregation methods, affinity propagation and k-means with k=5.**

it is ranked 81st and has a stable meaning, while our best model considered it one of the most changed and ranked it at 9.

We visualize the biggest clusters for *neutron* using PCA decomposition of BERT embeddings (Figure 1). There are two clearly distinctive clusters: cluster 36 in the bottom right corner, drawn with pink crosses, which consists only of instances from 1990s, and cluster 7 drawn with green dots in the top right corner, which consists only of instances from 1960s. A manual check reveals that the former cluster consists of sentences which mention *neutron stars*. Though neutron stars have been already discovered in 1960s they were probably less known[8] and are not represented in the corpus. In any case, a difference in a collocation frequency does not mean a semantic shift, since collocations often have a non-compositional meaning. Another similar example is a company called "Vector Security International" that appears only in 1990s time slice, which distorts semantic our calculations for the word *vector*. Our method could be improved by removing stable multiword expressions and named entities from the training set.

The latter distinctive cluster for *neutron*, consisting of word usages from 1960s, contains many sentences that have a certain pathetic style and elevated emotions, such as underlined in the examples below:

*throughout the last several decades the dramatic revelation of this new world of matter has been dominated by a most remarkable subatomic particle – the neutron .*

*the discovery of the neutron by sir james chadwick in 1939. marked a great step forward in understanding the basic nature of matter .*

The lack of such examples in 1990s might have a socio-cultural explanation or it could be a mere corpus artefact. In any case, this has nothing to do with semantic shift and demonstrates an ability of BERT to capture other aspects of language, including syntax and pragmatics.

[8]https://en.wikipedia.org/wiki/Neutron_star

## 6.2 Impact of Fine-tuning

Figure 2 shows the comparison of fine-tuning influence for two best clustering methods (affinity propagation, and k-means with k=5). Interestingly, a light fine-tuning (just for one epoch) decreases the performance of both methods (in terms of Spearman correlation) in comparison to no fine-tuning at all (zero epochs). After that, the length of fine-tuning until up to 5 epochs is linearly correlated with the performance increase.

Fine-tuning the model for five epochs appears optimal. After that, the performance for both methods starts decreasing, most likely because of over-fitting due to the reduced size of the fine-tuning dataset compared to the training data.

The impact of fine-tuning on the k-means clustering is stronger than on the affinity propagation. The difference between model's performance on 5 epochs is negligible. However, this effect holds only with k=5, other values of k do not demonstrate such a difference between original and fine-tuned models, as can be seen in Table 1.

## 6.3 Clustering

Results presented in Table 1 imply that most of the approaches for semantic change detection proposed in this work manage to outperform previous approaches by a large margin. We believe the differences in the numerical results should be primarily attributed to the differences in the methods, even though we can not draw a direct comparison to some of the approaches due to test set word removal and differences in the train corpora. We can however compare our results directly to the results published by [10] since they are also using BERT trained on the COHA corpus. Even more, their proposed clustering approaches are methodologically very similar to the approaches presented in this work, yet we manage to outperform their approach by a margin of about 35 percentage points when

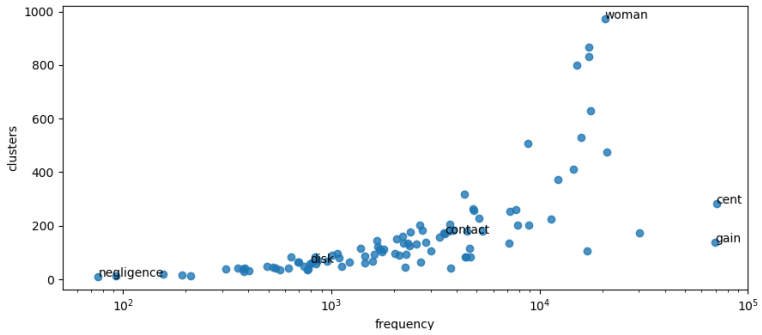



**Figure 3: Number of clusters found by affinity propagation and frequency of a word in the 1960s and 1990s in COHA.**

affinity propagation is used and by about 33 percentage points when k-means clustering[9], same as in [10], is used.

Unfortunately, [10] does not report a number of clusters that has been used, they only mention that the number of clusters has been optimised using the Silhouette scores. We can only speculate why their results are much lower than ours. The first hypothesis is connected with the usage of the Silhouette score, which might not be optimal for our goals. We compute the Silhouette score[10] for clusterings obtained by our methods. As can be seen in Table 1, the best Spearman correlation coefficient does not correspond to the best Silhouette score. Moreover, the Silhouette scores are quite close to zero.

The second hypothesis is connected with the difference in fine-tuning regimes employed in this research and the one conducted by [10]. We use domain adaptation fine-tuning, proving its efficiency for a certain number of epochs, for both k-means (except for a small number of clusters) and affinity propagation. However, [10] tried both diachronic fine-tuning (using the incremental fine-tuning technique first proposed by [18]) and domain-specific fine-tuning, but concluded that none led to an improvement in the results. As it was already speculated in [10], using both training regimes at the same time might lead to too extensive fine-tuning and therefore over-fitting. Further, a more thorough study on influence of incremental fine-tuning on contextual embeddings models (such as BERT) should perhaps be conducted, since the effects might differ from the ones observed for static embeddings models. Finally, the domain-specific fine-tuning is conducted only for 1 to 3 epochs, which might be too few to improve the results on some corpora.

The difference in performance between k-means and affinity propagation could be partially explained by the different number of clusters in the two approaches. Affinity propagation, which performs the best, outputs a huge amount of clusters—160 on average. The particular number of clusters found by affinity propagation for a word correlates strongly with the frequency of that word in the corpus with correlational coefficient $r = 0.875$, as is illustrated in Figure 3.

Thus, determining the optimal number of clusters for different words is not straightforward. We cannot claim that the clusters found by any of the methods we used can be interpreted as the different senses of a word or that they are even suitable for human interpretation. Most probably, affinity propagation captures subtle differences in word usages rather than global semantic shift. Nevertheless, it works better than k-means with smaller and more intuitive number of clusters, since word sense induction and semantic shift detection are not the same task.

Affinity propagation usually produces a skewed clustering, with a large number of small clusters containing only one or two data points, and can be used for outlier detection. K-means is not suitable for this task since it uses a random initialisation and if an outlier is not initially selected as a potential centroid it may never be found.

To justify this claim we conducted an additional experiment and run k-means clustering on fine-tuned embeddings using k=100 or number of instances minus one for less frequent words. As presented in Table 1, this resulted in Pearson and Spearman rank correlations of 0.315 and 0.337, respectively, which is worse than *any* other strategy we tried for fine-tuned embeddings, including averaging. At the same time, the Silhouette score for this insufficient model is almost equal to the Silhoutte score for the best model. Thus, the Silhouette score fails to discriminate between the best and the worst model.

## 7 FUTURE WORK

We plan to investigate how the clusters found by the methods in this work can be used to interpret the different usages of a word in a specific time slice. The initial experiments on this subject have already been conducted with the two-stage clustering, which removes the smallest clusters, containing one or two instances. Thus, it allows to focus on a smaller number of the most representative clusters, which might be more suitable for human interpretation even though it does not yield the best result. The initial check demonstrated that most of these clusters are interpretable, though some particular meaning can be spread among several clusters.

[9]Here we are referring to our best k-means configuration with five clusters and using a BERT model fine-tuned for five epochs.
[10]Using standard Scikit-learn implementation, https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

Our analysis hints that clustering BERT token embeddings for a word does not necessarily lead to sense-specific clusters. This conclusion is on par with [4]. Indeed, BERT ability do detect distinct word meanings has limitations. Thus, it would be interesting to extract only the semantic parts of the BERT embeddings to direct our analysis more towards word meaning and rather than word usage in general.

## ACKNOWLEDGMENTS

We are grateful to Andrey Kutuzov for valuable discussions during this paper preparation. We also thank Pr. Alexandre Allauzen from ESPCI - Université Paris Dauphine and Pr. Asanobu Kitamoto from National Institute of Informatics (Tokyo) for their advises. This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye) and 825153 (EMBEDDIA).

## REFERENCES

[1] Jean Aitchison. 2001. Language Change: Progress Or Decay? In *Cambridge Approaches to Linguistics*. Cambridge University Press, Cambridge.
[2] Domagoj Alagić, Jan Šnajder, and Sebastian Padó. 2018. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[3] Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598* (2019).
[4] Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda B. Viégas, and Martin Wattenberg. 2019. Visualizing and Measuring the Geometry of BERT. In *NeurIPS*.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. (2019), 4171–4186.
[6] Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 457–470.
[7] Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics* 4 (2016), 31–45.
[8] Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* 315, 5814 (2007), 972–976.
[9] Nabeel Gillani and Roger Levy. 2019. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*. 94–99.
[10] Mario Giulianelli. 2019. *Lexical Semantic Change Analysis with Contextualised Word Representations*. University of Amsterdam - Institute for logic, Language and computation.
[11] Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics*. 241–247.
[12] Kristina Gulordava and Marco Baroni. 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*. 67–71.
[13] Martin Hilpert and Stefan Th Gries. 2008. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24, 4 (2008), 385–401.
[14] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).
[15] Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3899–3908.
[16] Patrick Juola. 2003. The time course of language change. *Computers and the Humanities* 37, 1 (2003), 77–96.
[17] Nobuhiro Kaji and Hayato Kobayashi. 2017. Incremental Skip-gram Model with Negative Sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 363–371.
[18] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. 61–65.
[19] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*. 625–635.
[20] Andrey Kutuzov. 2020. Diachronic contextualized embeddings and semantic shifts. In *press*.
[21] Andrey Kutuzov, Elizaveta Kuzmenko, and Lidia Pivovarova. 2017. Clustering of Russian Adjective-Noun Constructions Using Word Embeddings. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. 3–13.
[22] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1384–1397.
[23] Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. In *LREC*.
[24] Hao Peng, Jianxin Li, Yangqiu Song, and Yaopeng Liu. 2017. Incrementally learning the hierarchical softmax function for neural language models. In *Thirty-First AAAI Conference on Artificial Intelligence*.
[25] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2227–2237.
[26] Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 474–484.
[27] Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv preprint arXiv:1811.06278* (2018).
[28] Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering* 24, 5 (2018), 649–676.
[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.
[30] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).
[31] Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2793–2807.

# Appendix C: Discovery team at SemEval 2020 - Task 1

## Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings not Always Better Than Static for Semantic Change Detection

**Matej Martinc** *
Jozef Stefan Institute, Slovenia
matej.martinc@ijs.si

**Syrielle Montariol**
Univ. Paris-Saclay, Societé Générale
LIMSI - CNRS, Univ. Paris-Sud, France
syrielle.montariol@limsi.fr

**Lidia Pivovarova**
University of Helsinki, Finland
lidia.pivovarova@helsinki.fi

**Elaine Zosa**
University of Helsinki, Finland
elaine.zosa@helsinki.fi

**Abstract**

The paper describes approaches used by the Discovery Team to solve SemEval-2020 Task 1— Unsupervised Lexical Semantic Change Detection. The proposed method is based on clustering of BERT contextual embeddings, followed by a comparison of cluster distributions across time. Best results were obtained by an ensemble of this method and static word2vec embeddings. According to the official results, our approach proved the best for Latin in the ranking subtask.

### 1 Introduction

Each word has a variety of senses and connotations, constantly evolving through usage in social interactions and changes in cultural and social practices. Identifying and understanding these changes is important for linguistic research and social analysis, since it allows detection of cultural and linguistic trends and possibly predict future changes. Detection of these changes can also be used for improvement of many Natural Language Processing (NLP) tasks, such as text classification and information retrieval.

The SemEval-2020 Task 1 — Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020) deals with detection of semantic change in temporal corpora containing texts from two distinct time periods in four languages: English, German, Latin and Swedish. The challenge defines two subtasks: Subtask 1 is a binary classification, i.e., to determine whether a word has changed or not; SubTask 2 aims at ranking a set of target words according to their rate of semantic change.

In this paper, we present the approaches used by the Discovery Team to tackle the two subtasks of SemEval-2020 Task 1. The Discovery Team qualified as 11th and 5th on Subtasks 1 and 2, respectively, and also proved the best for Latin language in Subtask 2. For quantifying and detecting semantic change, we leverage the transformer-based BERT model to generate contextualised embeddings for each word usage, and then aggregate these embeddings into meaningful time specific word representations by exploring different aggregation techniques, such as clustering (k-means and affinity propagation) and averaging. We also combine BERT-based representations with static word2vec embeddings. [1]

### 2 System Overview

#### 2.1 Word Representation

In order to derive meaningful temporal representations for each target word, we adapted the methodology proposed in Martinc et al. (2020a) to the multilingual setting of the SemEval-2020 Task 1. The core component of our approach is the usage of BERT (Bidirectional Encoder Representations from Transformers), a pretrained masked language model based on the transformer architecture (Devlin et al., 2019). We use specific models for each language[2], all with 12 attention layers and a hidden layer of size 768. German is the only language for which we use a cased model since most target words are nouns, which

*All authors contributed equally to this research.

[1]Code for the experiments is available under the MIT license at https://github.com/smontariol/Semeval2020-Task1.

[2]For English: bert-base-uncased model, for Swedish: bert-base-swedish-uncased (https://github.com/af-ai-center/SweBERT), for German: bert-base-german-cased (https://deepset.ai/german-bert), for Latin: bert-base-multilingual-uncased model.

are capitalized in German. The only model available for Latin is a multilingual BERT model trained on 100 languages, including Latin. For each language, the model is fine-tuned for five epochs on the task's corpus. This fine-tuning is unsupervised, i.e., a masked language model objective is used in the fine-tuning step (Devlin et al., 2019) in order to adapt each model to a specific corpus.

In the next step, the fine-tuned models are used to generate target-word embeddings. The corpus for each language is split into two time periods and the fine-tuned models are once again fed each sentence containing one or more target words from the subcorpus. A sentence embedding is generated for each of the input sentences by summing the last four encoder output layers of BERT. The resulting sentence embedding of size $sequence\ length \times embeddings\ size$ is finally split into tokens, to obtain a separate contextual embedding of size 768 for each token in the sentence. Note that byte-pair tokenization in some cases generates tokens that correspond to subparts of words. To generate embedding representations for the target words split into subparts, we concatenate embeddings for each byte pair token constituting a word. After this procedure, we have a contextual embedding representation for each target-word usage, together with the time period each word usage representation belongs to.

In addition to context-depended embeddings, we generate static word representations by training 300-dimensional Word2Vec model using Skip-gram negative sampling (SGNS) for each time slice. We align embeddings from the different time slices using the Orthogonal Procrustes (OP) method as in Hamilton et al. (2016). We also applied the pre- and post-processing steps recommended in Schlechtweg et al. (2019).

### 2.2 Measures of Semantic Change

We employ two methods for aggregating contextual embeddings: averaging and clustering. The methods were introduced and compared in our previous work (Martinc et al., 2020a; Martinc et al., 2020b).

**Averaging** is a simple aggregation approach where all target-word usage representations from a given time period are averaged. Then a quantitative estimate of semantic change for each target word is measured by computing the cosine distance between two averaged time-specific representations of the word.

**Clustering** of word usage representations results in sets of word usages, where each set is expected to correspond to a single word sense or a specific context. From the output of the clustering algorithms, we create two time-specific cluster distributions by normalizing the cluster counts within each period. Then the Jensen-Shannon divergence (JSD) between two time period-specific distributions is used to measure the semantic change. We used two clustering techniques, namely *affinity propagation* and *k-means*[3].

To obtain a quantitative estimate of semantic change for static embeddings, we measure the cosine distance between aligned embedding representations of the same word from two time slices.

### 2.3 Subtask 1: Binary Classification

In order to determine whether the word has changed or not, we experimented with two distinct methods, *thresholding using stopwords* and *identification of period-specific clusters*.

#### 2.3.1 Thresholding Using Stopwords

In the first method we try to find the best threshold in the ranking list based on the assumption that stopwords—words that are very frequent in a language but do not carry much meaningful information—undergo a low semantic change. We compute semantic change scores for a list of stopwords, employing the same procedure that was used for target words. Then, we compare stopword and target word score distributions in order to define a threshold below which a target word should be classified as unchanged.

For all languages except Latin, we create a list of stopwords by taking the words at the intersection of the `nltk` and `Spacy` stopword lists. For Latin, we use an external resource[4]. We keep only stopwords with more than 30 occurrences in each period; the number of stopwords per language is shown in Table 1. When the number of occurrences of a word is too high, we sample 5000 sentences per period for this word. As can be seen in Table 1, the mean JSD for stopwords is sensibly lower than the one for target words.

[3]We use the Scikit-learn implementations (`https://scikit-learn.org/stable/modules/clustering.html`) with default parameters, except for number of clusters for k-means, for which we tried several options.

[4]List of Latin stopwords: https://github.com/aurelberra/stopwords

| | | English | Latin | Swedish | German |
|---|---|---|---|---|---|
| Number of stopwords | | 109 | 334 | 78 | 142 |
| Mean JSD | stopwords | 0.181 | 0.210 | 0.355 | 0.328 |
| | targets | 0.239 | 0.264 | 0.460 | 0.384 |

Table 1: Number of stopwords used and average semantic change score (JSD) for target words and stopwords.

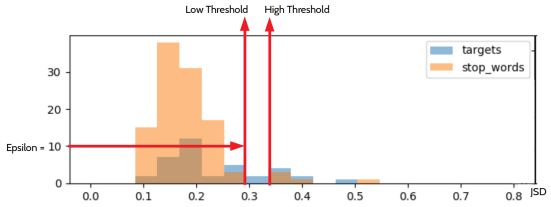



Figure 1: Distribution of semantic change scores in the English corpus: target words VS stopwords

| | aff-prop | avg | kmeans_5 | W2V | GS |
|---|---|---|---|---|---|
| aff-prop | 1 | | | | |
| averaging | 0.789 | 1 | | | |
| kmeans_5 | 0.815 | 0.811 | 1 | | |
| word2vec | 0.501 | 0.558 | 0.481 | 1 | |
| Gold Standard | 0.298 | 0.397 | 0.305 | 0.394 | 1 |

Table 2: Spearman correlation between the semantic change scores of various methods and the gold standard, averaged for all languages.

Though stopwords are more stable than most of the content words, they can still change their meaning due to the grammaticalisation processes. For example, the English stopword *hence* used to have a concrete physical meaning "from here" (e.g. "hence we go") but nowadays it is used only to connect two predicates. Since not all stopwords are stable, finding an appropriate threshold is not straightforward.

To define a threshold from the stopwords semantic change score distribution, we first divide it into 10 bins to derive a frequency distribution in a shape of a histogram with 10 columns, as exemplified for English in Figure 1. We take the threshold as the local maximum score of the bin in the histogram containing a number of words lower than an epsilon $\epsilon$. We exclude the first bin, which is composed of very stable words and can sometimes have a size smaller than $\epsilon$. The frequency limit $\epsilon$ used to select the threshold depends on the number of stopwords for each language: $\epsilon = 1/10 * \textit{number-of-stopwords}$. We compute two sets of thresholds: the leftmost and the rightmost points of the border bin, as shown in the Figure 1. The higher threshold is more conservative, meaning that less words are classified as changed.

#### 2.3.2 Identification of Period-Specific Clusters

The second method looks for concrete indications of semantic change, such as appearance or disappearance of a specific word sense. Target word clusters should to some extent resemble different word senses, allowing identification of target words that obtained or lost a meaning. If one of the clusters for a target word contains word occurrences from one time period and contains less or equal than $k$ (where k=2) word occurrences from another time period, we assume that this word has lost or gained a specific meaning.

Since clustering methods sometimes produce small-sized clusters, we consider only clusters bigger than a threshold, in order to focus on the "main" usages. Thus, for k-means we enforce a constraint that a cluster should contain at least 10 word occurrences to be considered in the analysis. For affinity propagation, we implement a dynamic threshold strategy: the threshold beyond which we consider a cluster is computed for each target word as twice its average cluster size.

### 2.4 Subtask 2: Ranking

For Subtask2, target words were ranked according to the semantic change scores described in Section 2.2, namely divergence between cluster distributions (JSD) and cosine distance. Additional steps were performed in some of our submissions to improve this basic approach: cluster filtering and ensembling.

#### 2.4.1 Cluster Filtering

We try several heuristics to filter out clusters that potentially contain noise and can distort comparison between time periods. The first idea is to remove smallest clusters (containing only one or two instances) whose appearance in a given time period is not significant. The second idea is to filter out sentences where a target word is used as a proper noun, e.g. as in the following example: *her daddy warn everyone that rose* ***lane_nn*** *be bring home a musician with long hair*.

Finally, we noticed that some clusters contain sentences that refer to specific events. For example, one of the clusters for *attack* contains sentences about terrorist attack in Israel and consists only of sentences from the latter time period, for the obvious reasons. The sentences in this cluster contain many named entities (NEs), e.g.: *hezbollah leader hassan fadlallah defend* ***attack_nn*** *on israeli civilian target civilian be a war crime*. We filter out clusters that contain too many NEs in some of our submissions, though this "radical" NE filtering may have drawbacks: e.g. one may argue that a "terrorist attack" is a new meaning of a word *attack* that was correctly distinguished by the clustering algorithm but then discarded by filtering.

In a real application NE recognition should be done on documents with preserved capitalization, preferably using a model trained specifically on historical documents. For the shared task we rely on out-of-the-box NLP pipelines.[5] Most of the tools are unable to recognize names in lowercased lemmatized text but POS-taggers are more reliable: e.g. SpaCy NE recognition model was unable to recognize lower-cased names even if SpaCy POS-tagger labeled the corresponding tokens as proper nouns.

We performed NE filtering as a post-processing step, to compensate for errors in NE recognition: a cluster is filtered out if the majority of the target word mentions are NEs. We filter out a cluster if at least 80% of the target word mentions are NE. For radical filtering, a cluster is filtered out if number of proper nouns is 5 times larger than a number of sentences.

#### 2.4.2 Ensembling

Different approaches for semantic change detection were ensembled by multiplying the semantic change scores produced by different methods for each target word. We experimented with different combinations of averaging, clustering and word2vec based methods in order to test the hypothesis that the synergy between contextualised and static embeddings improves the overall performance. Combinations of models that have too strong correlation (above 0.8) were discarded. Some correlations averaged for all languages can be found in table 2, though these values hide important disparities among languages.

## 3 Results

### 3.1 Subtask1

The results for the binary classification are shown in Table 3. We use BERT fine-tuned on the Semeval corpora for all submissions. The best official result was achieved by applying the stopword thresholding method to rankings obtained by measuring the JSD between affinity propagation cluster distributions. The method of identifying period-specific clusters worked competitively when conducted on k-means clusters but performed worse with affinity propagation, since the latter method usually produces a large number of clusters. Reducing the number of clusters by merging the closest clusters together increased the efficiency of the method.

The stopwords thresholding method seems to work best with higher thresholds, which classify less words as changed. But for all methods we face high discrepancies between languages. These can be clearly seen for the method of identifying period-specific clusters deployed on the affinity propagation clustering, which worked the best for Latin and the worst for all the other languages. Overall, the method of identifying period-specific clusters performed better for Swedish and Latin, while stopword thresholding worked better for English and German.

### 3.2 Subtask2

Results for SubTask 2 are presented in Table 4. The best official result was obtained by an ensemble of word2vec static embeddings and fine-tuned BERT contextual embeddings further improved with radical NE filtering as a postprocessing step—see row #11 in the table. The good performance of the method can be explained by the fact that the semantic change scores outputted using static embeddings and contextualised embeddings are not highly correlated, as shown in Table 2.

[5]We used SpaCy for English and German (`https://spacy.io/`), Polyglot for Swedish (`https://pypi.org/project/polyglot/`) and CLTK for Latin (`http://cltk.org/`).

| Model | Binary method | AVG | English | German | Latin | Swedish |
|---|---|---|---|---|---|---|
| k-means 5 | time-period specific clusters | 0.600 | 0.649 | 0.542 | 0.500 | **0.710** |
| aff-prop | time-period specific clusters, dynamic threshold | 0.496 | 0.568 | 0.458 | **0.700** | 0.258 |
| aff-prop, merging cluster | time-period specific clusters, dynamic threshold | 0.545 | 0.514 | 0.542 | 0.575 | 0.548 |
| aff-prop | stopwords, high threshold | 0.573 | 0.622 | 0.604 | 0.550 | 0.516 |
| aff-prop | stopwords, low threshold | 0.552 | **0.703** | 0.667 | 0.450 | 0.387 |
| ensemble: averaging + aff-prop | stopwords, low threshold | **0.621** | 0.568 | **0.688** | 0.550 | 0.677 |

Table 3: SubTask 1 results: accuracy. Submissions made during official evaluation phase are highlighted.

| | Input | Method | Post-Processing | AVG | English | German | Latin | Swedish |
|---|---|---|---|---|---|---|---|---|
| *Clustering* | | | | | | | | |
| 1 | pretrained BERT | aff-prop, JSD | - | 0.278 | 0.216 | 0.488 | **0.481** | -0.072 |
| 2 | fine-tuned BERT | aff-prop, JSD | - | 0.298 | 0.313 | 0.436 | 0.467 | -0.026 |
| 3 | fine-tuned BERT | aff-prop, JSD | small clusters | 0.302 | 0.327 | 0.440 | 0.472 | -0.030 |
| 4 | fine-tuned BERT | aff-prop, JSD | target NE | 0.300 | 0.328 | 0.426 | 0.467 | -0.023 |
| 5 | fine-tuned BERT | aff-prop, JSD | NE | 0.295 | **0.436** | 0.302 | 0.467 | -0.025 |
| 6 | fine-tuned BERT | aff-prop, JSD | NE, small clusters | 0.291 | 0.413 | 0.310 | 0.472 | -0.029 |
| 7 | fine-tune BERT | kmeans k=5, JSD | - | **0.320** | 0.189 | **0.528** | 0.324 | **0.238** |
| *Methods not using clustering* | | | | | | | | |
| 8 | fine-tune BERT | averaging, cosine dist | - | **0.397** | 0.315 | 0.565 | **0.496** | 0.212 |
| 9 | word2vec OP | cosine dist | (Schlechtweg et al., 2019) | 0.394 | **0.341** | **0.691** | 0.131 | **0.413** |
| *Ensembling* | | | | | | | | |
| 10 | aff-prop (#2) + w2v (#9) | distance multiplication | - | 0.417 | 0.357 | **0.642** | 0.366 | 0.303 |
| 11 | aff-prop (#2) + w2v (#9) | distance multiplication | NE, small clusters | 0.442 | **0.361** | 0.603 | **0.460** | 0.343 |
| 12 | aff-prop(#2), k-means (#7), averaging (#8), w2v (#9) | non-weighted multiplication | - | 0.403 | 0.279 | 0.607 | 0.451 | 0.276 |
| 13 | aff-prop(#2), k-means (#7), averaging (#8), w2v (#9) | weighted multiplication | - | **0.465** | 0.330 | 0.610 | 0.438 | **0.484** |

Table 4: SubTask 2 results: Spearman correlation with ground truth. Submissions made during official evaluation phase are highlighted.

Ensembling of four different methods—affinity propagation, K-means (k=5), averaging and word2vec OP—allows merging of all the information that they gather $(\#12)$. We improved this method by taking the correlation between the gold standard and each method as their respective ensembling weights (for each language) $(\#13)$. This yields better performance than the best of our submitted methods, though this is not an unsupervised approach and could only be done during the post-evaluation phase.

Cosine distance between averaged contextual embeddings performs much better than between word2vec representations for Latin but worse for other languages (rows #8 and #9). Affinity propagation clustering, which was the best in our previous study (Martinc et al., 2020a), did not perform well (rows #1 to #6), especially for Swedish, where it performed close to random. One explanation for this discrepancy could be sentence shuffling in the shared task corpora. BERT models cannot leverage the usual sequence of 512 tokens as a context in this setting but are limited to a number of tokens in the sentence. This could have a detrimental effect on the quality of their contextual embeddings. The results however do suggest that by averaging these embeddings a static embedding of good quality for each target token can be obtained.

Many approaches we tried improved performance only for English: BERT fine-tuning, affinity propagation clustering, NE filtering. This might be related to the fact that the corpora are lemmatized, and lemmatization has smaller effect on English, with its reduced morphology. Poor results on the Swedish corpus might be related to OCR-errors, leading to a large number of out-of-vocabulary tokens.

## 4 Conclusion

We have presented the approaches employed by the Discovery team to tackle SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). While our main method was based on clustering BERT contextualised embeddings, the best official result was obtained by combining this technique with a method for semantic change detection based on static word2vec embeddings.

The clustering-based methods are outperformed by embeddings averaging and word2vec-based method. The variety among languages is significant and the results averaged on all four corpora can be misleading.

## Acknowledgements

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye) and 825153 (EMBEDDIA).

## References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020a. Capturing evolution in word usage: Just add more clusters?

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020b. Leveraging contextual embeddings for detecting diachronic semantic shift. In *LREC*.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 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.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *To appear in SemEval@COLING2020*.

# Appendix D: Word clustering for historical newspaper analysis

## Word Clustering for Historical Newspapers Analysis

**Lidia Pivovarova** **Jani Marjanen** **Elaine Zosa**
University of Helsinki
firstname.lastname@helsinki.fi

**Abstract**

This paper is a part of a collaboration between computer scientists and historians aimed at development of novel methods for historical newspapers analysis. We present a case study of ideological terms ending with *-ism* suffix in nineteenth-century Finnish newspapers. We propose a two-step procedure to trace differences in word usages over time: training of diachronic embeddings on several time slices and when clustering embeddings of selected words together with their neighbours to obtain historical context. The obtained clusters turn out to be useful for historical studies. The paper also discusses specific difficulties related to development of historian-oriented tools.

### 1 Introduction

Big corpora of historical newspapers are now digitalized and available for automatic processing. Newspapers have for long been important sources of information for historians and social scientists but massive digitalization opens the possibility to use advanced statistical and NLP methods for historical newspapers. Even though news as a genre have been well-studied in NLP community, switching to historical news imposes additional difficulties for text processing. Automatically digitalized news archives contain much noise related to non-perfect OCR and article separation, as well as less standardised writing practices. Many NLP tools, such as POS-taggers and lemmatizers, are optimized to process modern texts and work less well on historical data. At the same time, historical news share most of the properties of the modern news data: they are biased, incomplete, controversial and apt to change over time.

If historical news are challenging for linguistic analysis, they are even harder for historical studies, since research questions historians are trying to answer are complex and lie far beyond fact discovery. Often they are interested in attitudes, stances, viewpoints, and discourse change in general. These tasks require development of novel methods and instruments that would be oriented specifically at historical research.

We present NewsEye—a research project aimed at development of novel tools and methods for analysis of historical newspapers[1]. The project is a collaboration between digital humanists and computer scientists funded by the European Union's Horizon 2020 research and innovation programme.

This paper focuses on a case study of ideological terms ending with *-ism* suffix—such as *liberalism*, *socialism*, or *conservatism*—in nineteenth century newspapers from Finland. These terms, known as isms, are condensed representations of complex notions that played an important role in political discourse in the nineteenth century (and long after that). Rhetorical usage of isms in historical text has been studied before (Kurunmäki and Marjanen, 2018b,a; Marjanen, 2018), though as far as we are aware this is the first attempt to apply statistical analysis to trace development of these terms in a diachronic newspaper archive.

Not all words ending with *-ism* are ideological. This suffix could be also used for medical terms and diseases (*rheumatism*), scientific terms (*magnetism*), personal traits (*cynicism*), artistic movements (*cubism*), religions (*baptism*) or political practices related to particular persons (*bonapartism*). It is not always possible to draw a strict line between ideologies and other categories.

[1] https://www.newseye.eu/



*Proceedings of the Workshop on Language Technology for Digital Historical Archives in conjuction with RANLP-2019*, pages 3–10,
Varna, Bulgaria, Sep 5, 2019.
http://doi.org/10.26615/978-954-452-059-5_002

Moreover, the ideological load of these terms might change over time.

We apply a corpus-based analysis to find out how the vocabulary of isms changed in nineteenth century Finnish newspapers and how usage of ideological isms is different from other words with *-ism* suffix. We try to implement a robust analysis procedure that would be applicable to other tasks with minimal human intervention. Our method consists of two main steps: first, we extract from the corpus *all* words with suffix *-ism*, second, we cluster these words and their semantic neighbours in an unsupervised fashion. This procedure does not require a human intervention other than interpretation of results and, consequently, is potentially applicable to other research questions.

## 2 Data

### 2.1 Corpora

Newspapers in Finland were published in two main languages—Finnish and Swedish. In the beginning of the nineteenth century the majority of newspapers were published in Swedish, though by the 1880s the Finnish and Swedish newspapers were printed in almost equal amount. The Finnish- and Swedish-language press had a different distribution of topics and exposed slightly different political outlook, though contemporaries often relied on newspapers in both languages (Engman, 2016). Another peculiarity of these data is a censorship accomplished by the Russian Empire government. The censorship was abandoned in 1905, which led to an outburst of socialistic rhetoric in the press, especially in the Finnish-language newspapers since they were more likely to have a rural or working-class background.

We use a digitalized collection of nineteenth-century Finnish newspapers freely available from the National Library of Finland (Pääkkönen et al., 2016). We use the full Swedish and Finnish data from 1820 to 1917, treating them as two separate corpora. Each corpus is split into five double-decades. The total amount of words in both corpora is presented in Table 1.

In Figure 1 we present relative frequencies for the selection of most frequent isms in our data. It can be seen that a proportion of isms are growing over time. The plots demonstrate some difference between the datasets: e.g. *patriotism* is much more frequent in the Swedish dataset.

| Time slice | Millions of words | |
|---|---|---|
| | FINNISH | SWEDISH |
| 1820-1839 | 1.3 | 25.5 |
| 1840-1859 | 10.3 | 77.9 |
| 1860-1879 | 90.6 | 326.7 |
| 1880-1899 | 805.3 | 966.9 |
| 1900-1917 | 2439.0 | 953.0 |
| **Total** | 3346.6 | 2355.2 |

Table 1: Corpus size by double decade.

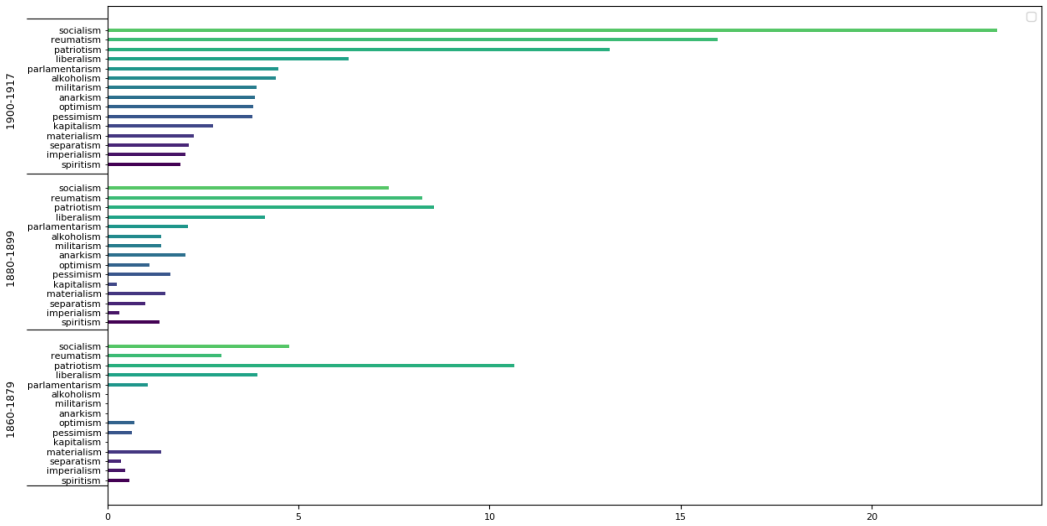
Both corpora are lowercased and lemmatized using LAS, an open-source language-analysis tool (Mäkelä, 2016).[2] LAS is a meta-analysis tool that provides a wrapper for many existing tools developed for specific tasks and languages. Though LAS supports multiple languages, most efforts were done to process Finnish data, including historical Finnish. The output for our Swedish data is more noisy. In particular, the Swedish LAS lemmatizer is unable to predict lemma for out-of-vocabulary words, e.g. *boulangismen* (definite form of 'boulangism'). Thus we applied the additional normalization and convert all words ending with *-ismen* or *-ismens* into *-ism* forms. For all other words we use the LAS output; implementation of proper Swedish lemmatization is beyond the scope of this paper.

## 3 Approach

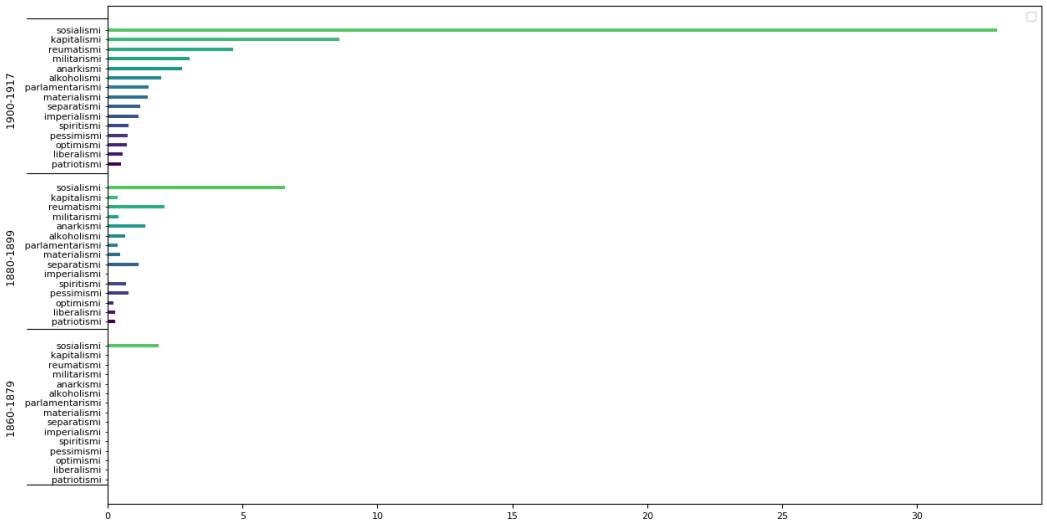
### 3.1 Diachronic embeddings

We train continuous embeddings (Mikolov et al., 2013) on each double-decade. We use Gensim Word2Vec implementation (Řehůřek and Sojka, 2010) using the Skip-gram model, with a vector dimensionality of 100, window size of 5 and a frequency threshold of 100—only lemmas that appear more than 100 times within a double decade are used for training. One hundred is an arbitrary and rather conservative threshold that ensures that each word in a model has reliable amount of context and embeddings are trustworthy. On the other hand, we lose some *isms* because they appear less than 100 times in a double-decade. For instance, *patriotism* and *liberalism* appear for the first time in the Swedish corpus in 1791 and 1820 respectively, but the corresponding vectors exist in our models starting from 1820-1839 and 1840-1859 respectively. The number of distinct isms in our models is presented in Table 2.

[2] https://github.com/jiemakel/las




(a) SWEDISH




(b) FINNISH

Figure 1: A selection of the most frequent words ending with suffix *-ism/ismi*. The x-axis presents relative frequency in items per million.

Since training word embeddings is a stochastic process, the particular values of vectors do not stay close across runs, though distances between words are quite stable. To ensure that embeddings are stable across time slices, we follow the approach proposed in (Kim et al., 2014): embeddings for $t + 1$ time slice are initialized with vectors built on $t$; then training continues using new data. The learning rate value is set to the end learning rate of the previous model, to prevent models from diverging rapidly. This approach has been previously used in (Hengchen et al., 2019) with slightly different data.

### 3.2 Clustering

We cluster word embeddings into semantically close groups using Affinity Propagation clustering technique (Frey and Dueck, 2007). The main advantages of Affinity Propagation are that it detects number of clusters automatically and is able to produce clusters of various sizes.

FINNISH

| Time slice | *ism* | *close* | *cluster* | *select* |
|---|---|---|---|---|
| 1820 - 1839 | 0 | - | - | - |
| 1840 - 1859 | 0 | - | - | - |
| 1860 - 1879 | 1 | 157 | 1 | 12 |
| 1880 - 1899 | 35 | 5977 | 20 | 442 |
| 1900 - 1917 | 119 | 8940 | 70 | 1543 |

SWEDISH

| Time slice | *ism* | *close* | *cluster* | *select* |
|---|---|---|---|---|
| 1820 - 1839 | 3 | 724 | 3 | 49 |
| 1840 - 1859 | 17 | 1845 | 12 | 211 |
| 1860 - 1879 | 61 | 5229 | 31 | 669 |
| 1880 - 1899 | 120 | 12233 | 54 | 1320 |
| 1900 - 1917 | 137 | 11858 | 56 | 1387 |

Table 2: Number of distinct words used on various steps of the algorithm: ***ism*** is a number of distinct words with suffix *-ism*, ***close*** is a number of words, which cosine similarity to at least one ism is higher than 0.5, ***cluster*** is a number of clusters that contain at least one ism, ***select*** is a number of words in these clusters.

Affinity Propagation has been previously used for various language analysis tasks, including collocation clustering into semantically related classes (Kutuzov et al., 2017) and unsupervised word sense induction (Alagić et al., 2018). Both papers pay special attention to fine-tuning of the algorithm and selection of hyper-parameters. We cannot tune the algorithm due to the lack of gold standard, which is typical for exploratory historical research. We use standard implementation from the Scikit-learn package (Pedregosa et al., 2011), with default parameters.

The procedure works as follows. In the data selection step we extract from the corpus all words with a cosine similarity of less than 0.5 to any *ism*. Then we perform clustering on this enriched dataset. Finally, the clusters are filtered so that only clusters that contain at least one *ism* word are presented for the qualitative analysis.

The number of words used on various steps of analysis is presented in Table 2. It can be seen from the table is that the number *isms* in the Finnish data is much smaller than for the Swedish data. In particular in the two double decades there are no Finnish ism above the frequency threshold. That could be partially explained by the smaller amount of Finnish newspapers but also by the difference between languages. The suffix *ismi* is not as productive in the Finnish language and used mostly with loan words, while Swedish more readily adopt *ism* suffix. In many cases Swedish words ending with *-ism* are translated into Finnish using native suffixes. For example, Swedish *katolicism* is translated into Finnish as *katolilaisuus*. In some cases, two words with same meaning but different endings existed in the same time period, e.g. *protestantismi* and *protestanttisuus* or *nationalismi* and *kansallisuusaate*.

It can be seen in the table that though 0.5 is an arbitrary threshold up to 90% of words selected using this threshold are filtered out after the clustering. The number of selected clusters is generally smaller than the number of words with suffix *ism* since isms tend to cluster together.

## 4 Results and Observations

One of the main difficulties for our work is a lack of gold standard annotations. We cannot know in advance how the words should be clustered, especially the most problematic ideological terms, which are the main objects of our study. However, we can make several common-sense assumptions on the expected outcome. For example, it would be reasonable to expect that disease names should not appear in the same cluster with philosophical concepts or that artistic movements should be clustered together. In this section we present several observations, starting with those that can be considered as "sanity checks" for the clustering.

### Rheumatism

In the nineteenth century rheumatism was often mentioned in the medical advertisements. Automatic advertisement filtering in historical news is not a trivial task since advertisements were less regulated, contained more text and looked similar to other articles. Moreover, such filtering is not always necessary since advertisements might provide researchers with valuable insights[3].

We use the entire corpora to build embeddings, and as a consequence *rheumatism* is one of the most frequent words with suffix *-ism* in our data, as can be seen in Figure 1 (for the Swedish data we sum up counts for spelling variants *reumatism* and *rheumatism*).

Table 3, which shows all clusters from our Finnish data that contain words related to rheuma-

[3]See for example a recent blog post analyzing gender stereotypes in the nineteenth century drug advertisements: https://www.newseye.eu/blog/news/british-drug-advertising-in-the-19th-century-through-the-prism-of-gender/

| 1880-1899 | 1900-1917 |
|---|---|
| *reumatismi* ‘rheumatism’<br>*luuvalo* ‘gout’<br>*luumalo* ‘gout’$_{ocr}$<br>*iskä* ‘?’ *latus* ‘?’<br>*liikavarvas* ‘callus’<br>*kihti* ‘gout’<br>*säilöstystauti* ‘canning disease’<br>*jalkahiki* ‘foot odor’<br>*kivuton* ‘painless’<br>*reumatillinen* ‘rheumatic’<br>*reumaatillinen* ‘rheumatic’ | *vähäverisyys* ‘anaemia’ *risatauti* ‘lymphadenitis’ *veripuute* ‘anaemia’ *heillou* ‘weakness?’$_{ocr}$<br>*nivelreumatismi* ‘arthritis’ *epämuodostuma* ‘deformity’ *kohju* ‘hernia’<br>*kroonillinen* ‘chronic’ *mahatauti* ‘gastroenteritis’ *mahakatarri* ‘gastritis’<br>*suolitauti* ‘salt deposits’ *riisitauti* ‘rickets’ *hermovaiva* ‘nerve ailment’<br>*verenvähyys* ‘anaemia’ *ruumisvika* ‘body problem’ *veritauti* ‘blood disease’<br>*lihavuus* ‘obesity’ *kaljupäisyys* ‘boldness’ *verettömyydä* ‘verettömyydä’<br>*heikkohermoisuus* ‘neurasthenia’ *lihanen* ‘obese’ *sukupuoli-* ‘sex/gender’$_{ocr}$<br>*sappitauti* ‘biliary disease’ *heitlous* ‘weakness’$_{ocr}$ *selkäydintauti* ‘spinal cord disease’<br>*hermoheikkous* ‘neurasthenia’ *ruokasulatushäiriö* ‘digestion problem’<br>*kalvetustauti* ‘anaemia’ *vinous* ‘skewness’ *tautitila* ‘disease place’<br>*vähäverinen* ‘anaemic’ *epämuodostua* ‘to deform’ *hermosairaus* ‘neuropathy’ |


| |
|---|
| *reumatismi* ‘rheumatism’ *hiustauti* ‘hair disease’ *jäsensärky* ‘limb ache’<br>*hermo* ‘nerve’ *oxygeno* ‘?’ *vatsakatar* ‘gastritis’ *umpitauti* ‘constipation’<br>*nuha* ‘rhinitis’ *hermotautinen* ‘neurotic’ *topioli* ‘?’ *kurkkukatarri* ‘pharyngitis’<br>*parannuskeino* ‘remedy’ *hoitokeino* ‘cure’ *spirosiini* ‘spirosin’ *lazarol* ‘lazarol’<br>*lääkitä* ‘to medicate’ *kotilääke* ‘home medicine’ *reumaattinen* ‘rheumatic’<br>*hammastauti* ‘tooth disease’ *rautaliuos* ‘iron care’ *jäsenkolotus* ‘limb ache’<br>*leini* ‘rheumatism’ *linjamentti* ‘ointment’ *parannusaine* ‘betterment’ *vilustuminen* ‘cold’<br>*luuvalo* ‘gout’ *latsaro* ‘?’ *hengityselimettauti* ‘respiratory disease’ |

Table 3: Clusters containing Finnish words related to rheumatism. Original words are presented in italics together with English translations in quotes. *ocr* means the word is incorrectly spelled due to OCR errors; “?” means “impossible to translate”—these are mostly fragments of words appearing due to OCR errors. Bottom left: an advertisement of a rheumatism medicine from *Hufvudstadsbladet*, 01.03.1912, no. 59, p. 15

tism. It can be seen that *rheumatism* does not interfere with other isms: the clusters entirely consist of words related to drugs, medical procedures, diseases and other physical conditions, such as baldness or obesity. In that sense clusters are rather precise and justify our algorithmic decisions.

On the other hand, cluster may be too fine-grained for our needs. In the 1900-1917 double-decade there are two clusters with similar meaning: one related to *reumatismi* ‘rheumatism’, another to *nivelreumatismi* ‘(rheumatoid) arthritis’. Very similar results were obtained on the Swedish data: *reumatism* ‘rheumatism’ and *ledgngsreumatism* ‘arthritis’ are split into different clusters even though spelling variants *rheumatism* and *reumatism* are clustered together.

We suggest that the fine-grained clustering does not as such reflect semantic differences, but the differences in distribution come from slightly different uses in the newspapers. While there are similarities it seems that rheumatism appears more often in medical advertising whereas the arthritis seems to be more likely to appear in text content with a more ambitious take on educating the public about medical issues.

### Spiritism

In Table 4 we present clusters obtained from Swedish data that contain the word *spiritism*. The cluster for the 1860-1879 double decade contains a few words related to this popular practice such as *pressensé* and *kabal* though most of its content are names of famous scientists and writers. This might be an error: some of the names might be a person that were discussed in the context of spiritism (as objects to spiritism or as scientific authorities), e.g. Aristotle or Galileo, and others are words that are similar to these names. In other words, *spiritism* might be an outlier in this cluster.

It might also be the case that spiritism was sometimes used as ’spiritualism’ and Darwin and the others were discussed in this context. This would require a further analysis.

The clusters for the latter double-decades do not expose such problems and consist mostly of words clearly related to spiritism including some very specific terms, such as *transmigration*, and more general esoteric concepts, such *theosophy* or *freemasonry*. The 1880-1899 cluster might also reflect a contemporary discussion on relations between science and mysticism, since it contains such isms as *positivism* or *darwinism*.

### Separatism

Separatism is a more tricky concept, which undergo a noticeable usage change in our datasets as can be seen in Table 5, where we present clusters for Swedish *separatism*.

| 1860-1879 | 1880-1899 | 1900-1917 |
|---|---|---|
| *spiritism* 'spiritism'<br>*pressensé* 'presence' (Fr)<br>*pater* 'pater' *voltaire* 'Voltaire'<br>*darwin* 'Darwin' *renan* 'Renan'<br>*zola* 'Zola' *newton* 'Newton'<br>*balzac* 'Balzac' *michelet* 'Michelet'<br>*galilei* 'Galileo' *corneille* 'Corneille'<br>*aristoteles* 'Aristotle' *kabal* 'cabal'<br>*oppert* 'Oppert' *rousseau* 'Rousseau'<br>*proudhon* 'Proudhon' *zolas* 'Zola'<br>*quand* 'when' (Fr) *loyson* 'Loyson' | *spiritism* 'spiritism' *teosofi* 'theosophy'<br>*frimureri* 'freemasonry' *feder* '?'<br>*mysterium* 'mystery' *spiritualism* 'spiritualism'<br>*darwinism* 'darwinism' *positivism* 'positivism'<br>*buddism* 'Buddhism' *darvinism* 'darvinism'<br>*vegetarianism* 'vegetarianism' *astrologi* 'astrology'<br>*teosofisk* 'theosophic' *bibelkritik* 'Bible criticism'<br>*metafysik* 'metaphysics' *teosofien* 'theosophy'<br>*darvin* 'Darvin' *darvins* 'Darvin'<br>*utvecklingslära* 'evolution' *malthus* 'Malthus'<br>*självandring* 'transmigration' | *spiritism* 'spiritism' *hypnotism* 'hypnotism'<br>*andevärld* 'spirit world' *teosofisk* 'theosophic'<br>*spiritistisk* 'spiritualistic' *telepati* 'telepathy'<br>*självandring* 'transmigration'<br>*trolleri* 'magic' *journalism* 'journalism'<br>*ockult* 'occult' *astrologisk* 'astrological'<br>*astrologi* 'astrology' *frimureri* 'freemasonry'<br>*gondiagnos* 'eye diagnosis' *alkemi* 'alchemy'<br>*clairvoyance* 'clairvoyance'(Fr)<br>*tankeläsning* 'mind reading'<br>*tungomlstalande* 'tongues' |

Table 4: Clusters containing Swedish word *spiritism*.

| 1860-1879 | 1880-1899 | 1900-1917 |
|---|---|---|
| *separatism* 'separatism'<br>*mysticism* 'mysticism' *naturalism* 'naturalism'<br>*darwinism* 'darwinism' *moral* 'morality'<br>*tidsanda* 'zeitgeist' *krass* 'crass' *utopi* 'utopia'<br>*materialistisk* 'materialistic' *otro* 'incredible'<br>*rationalistisk* 'rationalistic' *wantro* '?'<br>*menniskonaturen* 'human nature' *tidehvarfvets* '?'<br>*materialism* 'materialism' *materialist* 'materialistic'<br>*konservatism* 'conservatism'<br>*idealism* 'idealism' *rationalism* 'rationalism'<br>*negation* 'negation' *abstraktion* 'abstraction'<br>*idealistisk* 'idealistic' | *separatism* 'separatism' *rent* '?'<br>*finskhet* 'Finnishness' *fennomanins* 'Fennomania'<br>*fennomani* 'Fennomania' *svenskhet* 'Swedishness'<br>*fennomanin* 'Fennomania' *vikingaparti* 'Viking party'<br>*språkpolitik* 'language policy' *publicistisk* 'publishing'<br>*partiagitation* 'party agitation' *partiyra* '?'<br>*partifanatism* 'party fanaticism'<br>*språkgräl* 'language quarrel'<br>*språkfanatism* 'language fanaticism'<br>*språkfråga* 'language question'<br>*spräkfrägan* 'language question'<br>*ljusskygghet* 'photophobia' | *separatism* 'separatism' *riksid* 'national idea'$_{ocr}$<br>*statsid* 'state idea'$_{ocr}$ *rikspolitik* 'national policy'<br>*bourgeoisins* 'bourgeoisie' *byråkratien* 'bureaucracy'<br>*samhällsopinion* 'social opinion'<br>*sträfvandenas* '?' *rikskomplex* 'national complex'<br>*nationalitet-* 'national'$_{ocr}$ *santryska* 'true Russian'<br>*ämbetsmannavälde* 'officialdom'<br>*gränsmärke* 'borderline' *gränsmark* 'borderline'$_{ocr}$<br>*riksenhet* 'national assembly'<br>*samhällskraft* 'social force' *statlighet* 'statehood'<br>*frihetssträvande* 'freedom-aspiring' *wäldets* '?'<br>*riksmakt* 'national power' *självhärskarmakten* '?' |

Table 5: Swedish clusters containing word *separatism*

| 1880-1899 | 1900-1917 |
|---|---|
| *separatismi* 'separatism' *ruotsi-kiihkoinen* 'Svekoman' *ruotsinmielinen* 'Swedish-minded'<br>*ruotsalaisuus* 'Swedishness' *viikinki* 'Viking' *ruotsi-mielinen* 'Swedish-minded'<br>*fennomaani* 'Fennoman' *epäkansallinen* 'anti-national' *viikingit* 'Vikings'<br>*separatisti* 'separatist' *ruotsikko* 'Swedish'(person) *miikinki* 'Viking'$_{ocr}$ *pöppö* '?'<br>*miikingit* 'Vikings'$_{ocr}$ *suomimielinen* 'Finnish-minded' *ruotsi-mielisyys* 'Swedish-mindedness'<br>*wiitinki* 'Viking'$_{ocr}$ *wiilinki* 'Viking'$_{ocr}$ *miitinki* 'Viking'$_{ocr}$ *ruotsimielinen* 'Swedish-minded'<br>*suomi-kiihkoinen* 'Fennoman' *fennoman* 'Fennoman' *henkiheimolainen* 'soul mate'<br>*dagbladilainen* 'member of the Dagblad circle' *miiking* 'Viking'$_{ocr}$ *fennomani* 'Fennoman'<br>*wiiking* 'Viking'$_{ocr}$ *fennomaaninen* 'Fennoman' *ruotsikiihkoisuus* 'Svekomania'<br>*wiilinli* 'Viking'$_{ocr}$ *miikinkilehti* 'Vikings' newspaper'$_{ocr}$ *suomenmielinen* 'Finnish-minded'$_{ocr}$<br>*miikinkiläinen* 'Vikingish'$_{ocr}$ *ruolsinmielinen* 'Swedish-minded' *ruotsiliihloinen* 'Svekoman'$_{ocr}$<br>*herranenluokka* '?' *miikingilehti* 'Vikings' newspaper'$_{ocr}$ *epälansallinen* 'anti-national'$_{ocr}$ | *separatismi* 'separatism'<br>*nationalismi* 'nationalism' *natsionalismi* 'nationalism'<br>*opportunismi* 'opportunism' *natfionalismi* 'nationalism'$_{ocr}$<br>*eristäytyminen* 'isolation' *kansalliskiihko* 'nationalism'<br>*intelligens* 'intelligence' *länsieurooppalainen* 'Western-European'<br>*rotutaistelu* 'race fight' *vapaamielisyy* 'liberalism'$_{ocr}$<br>*sanomalehdistö!* 'press' *antipatia* 'antipathy'<br>*kansallinenviha* 'national anger' *kiihkokansallisuus* 'nationalism'<br>*eristäytyä* 'self-isolate' *liittolaisuus* 'alliance'<br>*vihamieli-syy* 'hostility'$_{ocr}$ *kansallinenylpeys* 'national pride'<br>*kielipolitiikka* 'language policy'<br>*kansallinenliike* 'national movement' |

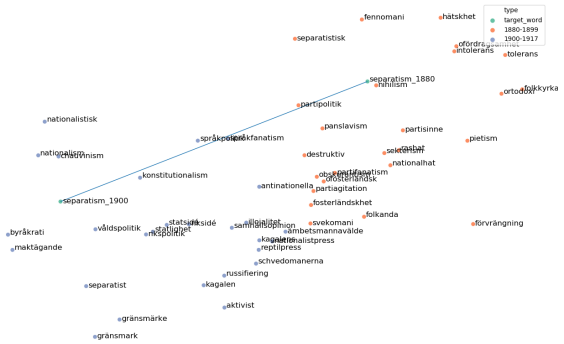
Table 6: Finnish clusters containing word *separatismi*

Most of the words in the 1860-1879 cluster are religious, philosophical or scientific notions, thus we can assume that the cluster presents a religious context of *separatism*. The 1880-1899 cluster contains completely different set of words, including reference to specific political entities, such as Fennomans movement and contains rather emotional expressions, such as *agitation* or *fanaticism*. These words are related to a contemporary discussion about national identity and national language. The 1900-1917 cluster is again different from the previous two and contains more general political lexis. Thus, we can suggest that at the beginning the notion of separatism had mostly religious meaning, when it was adopted by a limited number of liberals and finally spread into a more general political discourse.

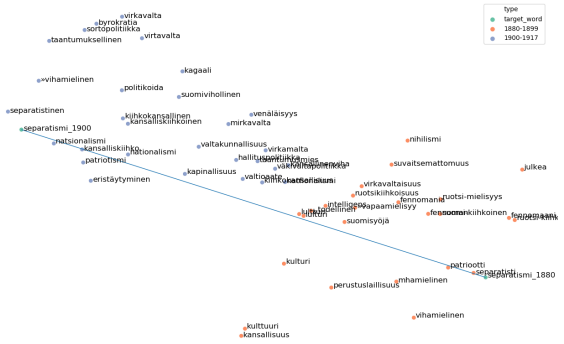
The Finnish clusters for *separatismi*, presented in Table 6, are quite similar to Swedish. The main difference is that in the 1860-1879 the word is mentioned less than 100 times and as a consequence excluded from our models. But the 1880-1899 and 1900-1917 Finnish clusters follow the same pattern: the former contains quite specific references, while the latter consists of more general political words.

The change in the distribution of *separatism* seems to be related to a change in the dominant context in which it was discussed (from religious context to a political context). This also entails some degree of semantic change.

This contextual and semantic shift could be to some extent visible from changes in the nearest neighbours of *separatism* presented in Figure 2a. However, nearest neighbours produce a more vague overview: for example, religious isms, such as *pietism*, are presented among nearest neighbours of *separatism* in 1860-1879. Similarly, the overlap between Finnish clusters, shown in Table 6, and nearest neighbours of *separatismi*, presented in Figure 6 is very limited.




(a) SWEDISH




(b) FINNISH

Figure 2: tSNE plot word *separatism* and its nearest neighbours across time slices.

This can be explained by the nature of the clustering procedure: each word can be among the nearest neighbours for any number of other words while Affinity Propagation assign a word to exactly one cluster so that *socialism* and *katolicism* are separated in clusters of their own. The difference between outputs demonstrates an added value of the clustering, which selects only one word split among many possibilities provided by embeddings. At the same time, this also means loss of information, especially for polysemous words.

## 5 Conclusion and Further Work

We presented our ongoing work aimed at the implementation of tools facilitating historical studies of newspaper archives. We proposed an unsupervised procedure to trace differences in word usage over time. The procedure consists of two major steps: training of diachronic embeddings and then clustering embeddings of selected words together with their neighbours to obtain historical context.

In this paper we applied this procedure to a group of words ending with suffix *-ism*. The method allowed us to distinguish ideological terms, such as *socialism* from other words with the same suffix, such as disease names or scientific terms. This promising result suggests that it is worthy to further elaborate the proposed method.

At this stage of the work we are unable to draw any clear conclusions related to usage of isms in the nineteenth century in Finland. Clusters that contain ideological words are the most problematic for the interpretation, which is not surprising given complex nature of the underlying concepts.

Nevertheless, we consider the obtained clusters useful for historical studies since they provide a researcher with a condensed representation of word usages in a large corpus. This is a novel way to look at historical data, which might be especially useful in combination with other tools such as named entity recognition or topic modelling.

Further improvements of the method should include both parts, namely embeddings and clustering. We plan to try building *continuous* word embeddings (Dubossarsky et al., 2019; Gillani and Levy, 2019; Rosenfeld and Erk, 2018; Yao et al., 2018) that would allow us to investigate gradual semantic shifts rather than split data into discrete time slices. Improvement of clustering might include fine-tuning of the algorithm parameters, though this is quite hard to do without manually annotated data. Thus, our main focus would be in finding other applications for the proposed procedure that would be meaningful from a historical research point of view and easily assessed at the same time.

We will also continue development of complex instruments for historical news analysis that would utilize clustering techniques together with other automatic text analysis methods.

## Acknowledgements

We are grateful to Simon Hengchen and Mark Granroth-Wilding for the help with data preparation. This work has been supported by the European Unions Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

## References


Domagoj Alagić, Jan Šnajder, and Sebastian Padó. 2018. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Max Engman. 2016. *Språkfrågan: Finlandssvenskhetens uppkomst 1812-1922*. Svenska litteratursällskapet i Finland.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315(5814):972–976.

Nabeel Gillani and Roger Levy. 2019. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *NAACL HLT 2019*. page 94.

Simon Hengchen, Ruben Ros, and Jani Marjanen. 2019. A data-driven approach to the changing vocabulary of the nation in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In *In Proceedings of the Digital Humanities (DH) conference 2019, Utrecht, The Netherlands*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *ACL 2014* page 61.

Jussi Kurunmäki and Jani Marjanen. 2018a. Isms, ideologies and setting the agenda for public debate. *Journal of Political Ideologies* 23(3):256–282. https://doi.org/10.1080/13569317.2018.1502941.

Jussi Kurunmäki and Jani Marjanen. 2018b. A rhetorical view of isms: An introduction. *Journal of Political Ideologies* 23(3):241–255.

Andrey Kutuzov, Elizaveta Kuzmenko, and Lidia Pivovarova. 2017. Clustering of Russian adjective-noun constructions using word embeddings. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. pages 3–13.

Eetu Mäkelä. 2016. LAS: an integrated language analysis tool for multiple languages. *The Journal of Open Source Software* 1.

Jani Marjanen. 2018. Ism concepts in science and politics. *Contributions to the History of Concepts* 13(1).

Tomas Mikolov, Kai Chen, Greg S Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *NIPS*.

Tuula Pääkkönen, Jukka Kervinen, Asko Nivala, Kimmo Kettunen, and Eetu Mäkelä. 2016. Exporting Finnish digitized historical newspaper contents for offline use. *D-Lib Magazine* 22(7/8).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pages 474–484.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *The 11th ACM International Conference on Web Search and Data Mining*.

# Appendix E: Clustering ideological terms in historical newspaper data with diachronic word embeddings

## Clustering Ideological Terms in Historical Newspaper Data with Diachronic Word Embeddings

Jani Marjanen[1], Lidia Pivovarova[1], Elaine Zosa[1], and Jussi Kurunmäki[2]

[1] University of Helsinki, Helsinki, Finland
firstname.lastname@helsinki.fi
[2] University of Tampere, Tampere, Finland
firstname.lastname@tuni.fi

**Abstract**

During the course of the nineteenth century, ideological language mostly expressed through *isms* such as liberalism, socialism or conservatism, entered the lexicon in most European languages. Previous research has based on reading key texts claimed that the suffix *ism* was introduced to new linguistic domains during the period up to WWI, many of which do not relate to ideology. This paper uses a data-driven way to study the emergence of *isms* in nineteenth-century Finnish newspapers and uses word embeddings to cluster them and to trace their thematic expansion in the period. As such, the study provides a quantitatively sound way of tracking how *isms* relate to ideological language and more generally contributes to the understanding of the development of political language in Finland.

CEUR-WS.org/Vol-2461/paper_4.pdf

### 1 Introduction: A data-driven perspective on *isms* and ideology

Words ending in the suffix *-ism* are terms that reduce complex figures of thought under one simple heading. As such they are emotionally evocative, communication-wise effective, and also contested in meaning [24]. *Isms* are commonly associated with ideologies in modern political language, but not all *isms* are ideologies and vice versa [5]. In fact the relationship between *isms* and notions of ideology have changed historically and has varied depending on cultural context [11].

While there are works that study *isms* from a long-term perspective [9], their applicability as analytical tools for historiography [2], their rhetorical appeal [12], or their cultural transferability and special place in Chinese political language [24], there are no quantitative studies that try to describe how central political and social words with the suffix *-ism* have changed over time and how they relate to one another. This paper takes a step in that direction by approaching long-term data set of historical newspaper text from Finland and using word embeddings to analyze how *isms* related to one another in the long nineteenth century.

In the recent year distributional semantics methods have been applied to assess how lexical and semantic change manifests itself in historical corpora [14, 25]. They more or less rely on the so-called distributional hypothesis that in different variants posits that words' similar distribution in context indicates a similarity also in meaning [23]. While this paper does not assume one to one correlation between distribution and semantics, it uses word embeddings

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

to cluster different *isms* and words close to them in the distributional space over time. In doing so we assume that the distributions allow for clustering *isms* either according to semantic similarity or similarity in rhetorical tropes or pragmatics.

In this paper we do not use methods based on comparing a word vector from one time slice to a vector for the same word in another time slice, since those methods are aimed at finding radical changes in word senses, such as discussing new sense acquired by words like *gay* or *computer* in the twentieth century. Words that we are primarily interested in this study did not undergo such radical transformations—e.g. *patriotism* meant more or less 'love for one's country' throughout the whole nineteenth century—though context and valuation of its usage changed. Instead, we apply clustering of word vectors and demonstrate that word clusters changed as the context of the *ism* vocabulary was expanded over time.

Clustering *isms* over a long period of time in a data-driven way poses a number of methodological problems, which requires testing and exploration. The potential benefit of doing this lies in producing a statistically robust image of how *isms* developed. Earlier studies have argued that *isms* transformed from the religious sphere, to the political and ideological sphere in the late eighteenth century and early nineteenth century with pivotal *isms* such as patriotism, liberalism and socialism transforming the field. The field of *isms* further expanded in the late nineteenth century with new *isms* in philosophy, science and arts appeared [9, 11]. A data-driven clustering currently already shows how the vocabulary of *isms* indeed expanded over the nineteenth century and how the political *isms* do cluster quite heavily, whereas medical words ending with the same suffix, such as the very common word *rheumatism*, are definitely kept separate from any ideological debate revolving around *ism* words. Our analysis also suggests that with changes in political context key *isms* were clustered differently based on the political situation they described. This change is partly about changes in semantics, but not only. For instance, an *ism* like 'socialism' did have a remarkable semantic continuity throughout the nineteenth century, but what it meant for newspapers to write about socialism changed when socialism had been associated more with radicalized political events. Contestation regarding socialism had much to do with potential radical futures associated with it.

## 2 Research questions, methods and data

### 2.1 Research questions

This paper studies *isms* as particularly laden keywords in societal discourse in Finland in the long nineteenth century. We address the following research questions:

- How did the vocabulary of *isms* expand in the period?
- Which *isms* appear as similar based on their embeddings?
- Are there interesting continuities in the enriched clustering that takes into account nearest neighbors of the *isms*?

Finally, we shortly discuss the differences in Finnish-language and Swedish-language discourse in Finland when looked upon through *isms*.

### 2.2 Data

We use a digitalized collection of nineteenth-century Finnish newspapers freely available from the National Library of Finland [19]. Though the archive contains newspapers starting from

Table 1: Corpus size by double decade.

| Time slice | Millions of words | |
|---|---|---|
| | FINNISH | SWEDISH |
| 1820–1839 | 1.3 | 25.5 |
| 1840–1859 | 10.3 | 77.9 |
| 1860–1879 | 90.6 | 326.7 |
| 1880–1899 | 805.3 | 966.9 |
| 1900–1917 | 2439.0 | 953.0 |
| **Total** | 3346.6 | 2355.2 |

1770s, the earlier time periods do not have enough data for the automatic analysis we apply in this paper. Thus, we use data from 1820 to 1917. The collection contains newspapers in the Russian, German, Swedish and Finnish languages, with the latter two as the dominant languages. In our analysis, these dominant languages are treated as two separate corpora even though contemporaries often relied on newspapers in both languages [4]. The total amount of words in both corpora is presented in Table 1.

Both corpora are lowercased and lemmatized using LAS, an open-source language-analysis tool [15]. LAS is a meta-analysis tool that provides a wrapper for many existing tools developed for specific tasks and languages. Though LAS supports multiple languages, most efforts were done to process Finnish data, including historical Finnish. The output for our Swedish data is more noisy. In particular, the Swedish LAS lemmatizer is unable to predict lemma for out-of-vocabulary words, e.g. *boulangismen* (definite form of 'boulangism'). Thus we applied the additional normalization and convert all words ending with *-ismen* or *-ismens* into *-ism* forms. For all other words we use the LAS output; implementation of proper Swedish lemmatization is beyond the scope of this paper.

## 2.3 Diachronic embeddings

To trace semantic shifts in word meanings we split a lemmatized corpus into double decades (1820–1839, 1840–1859, and so on until 1900–1917) and train continuous embeddings [18] on each time slice. We use the Gensim Word2Vec implementation [21] using the Skip-gram model, with a vector dimensionality of 100, window size 5 and a frequency threshold of 100—only lemmas that appear more than 100 times within a double decade are used for training. That way we try to ensure that each word in a model has reliable amount of context and the embeddings are trustworthy. However, we lose some *isms* because they appear less than 100 times in a double-decade. For example, the Finnish word *feminismi* was mentioned 91 times between 1900 and 1917 and was excluded from our analysis, while its Swedish counterpart was mentioned 242 times and is visible in our results. Our models allow us to detect when a word became frequent, in what context it was used and what is the difference between Swedish and Finnish contexts. They do not allow, however, to check when the word appeared for the first time and comparison of word distributions between languages is not fully reliable for less frequent words.

Since training word embeddings is a stochastic process, the particular values of vectors do not stay close across runs, though distances between words are quite stable. To ensure that embeddings are stable across time slices, we follow the approach proposed in [10]: embeddings

https://github.com/jiemakel/las

for $t+1$ time slice are initialized with vectors built on $t$; then training continues using new data. The learning rate value is set to the end learning rate of the previous model, to prevent models from diverging rapidly. This approach has been previously used in [8] with slightly different data.

## 2.4 Clustering

To investigate the expansion of the vocabulary of *isms* we cluster words into semantically close groups. Since our task is mostly exploratory and the number of clusters cannot be known in advance we apply the Affinity Propagation clustering technique [6]. The method splits all datapoints into *exemplars*, i.e. cluster representative tokens, and *instances*, i.e. other members of clusters. At the initial step all datapoints present a cluster of their own. Then for each instance-representative pair a likelihood for an instance to be represented by an exemplar is computed by taking into account all other instances of the exemplar and all other available exemplars for the instance. This computation is repeated until convergence; if an exemplar has no instances it is dismissed. We use standard implementation from Scikit-learn package [20], with default parameters.

Affinity Propagation has been previously used for various language analysis tasks, including collocation clustering into semantically related classes [13] and unsupervised word sense induction [1]. The main advantages of the method are that it detects the number of clusters automatically and is able to produce clusters of various size. As a side effect it returns exemplars, i.e. cluster representatives, which are not necessary equal to the geometric centre of the cluster.

The main drawback of the Affinity Propagation is pairwise computations. The method is quadratic in time and memory and cannot be applied to large datasets, such as a whole corpus vocabularly. Thus, data selection is an unavoidable step. In this paper we use Affinity Propagation in two experiments.

In the first experiment, we extract from the corpus all *ism* words. i.e. words that end with *-ism* in Swedish and *-ismi* in Finnish and cluster only this set of words. The extraction allows us to identify how close these words to each other given other *isms* in the corpus.

In the second experiment, we try to put *isms* into a richer context and trace other words associated with them in the respective double-decades. We extract from the corpus all words which have a cosine similarity to any *ism* that is less than 0.5. Then we perform clustering on this enriched dataset. Finally, the clusters are filtered so that only clusters that contain at least one *ism* word are presented for the qualitative analysis. An output of this procedure is different comparing to the first experiment, i.e. words that clustered together in the *ism*-only clustering can break up into different enriched clusters, since in the latter setting they have more exemplar options.

Clustering is performed separately for each time slice. To link clusters across time we perform visualization with Sankey charts. In the Sankey diagram, clusters from time slice $t$ are linked to clusters in time slice $t+1$ based on the number of words they have in common.

The magnitude of the link is the sum of the word frequencies (from the source cluster, that is the cluster from time slice $t$) of the common words between the connected clusters.

We exclude from the list words that are shorter than 5 characters for Swedish and 6 characters for Finnish. This is to filter out obvious OCR bugs such as *ism*, *tism*, *rism*, etc. Though the words 'ism' and 'ismi' exist in the Swedish and Finnish languages, they are very uncommon in nineteenth-century press.

# 3 Results

## 3.1 Swedish and Finnish clusters

As expected, Finnish-language and Swedish-language *isms* cluster differently in terms of timing and themes that are present. There are three main reasons for this:

- Swedish-language press in Finland developed earlier and included more abstract content earlier in the century, whereas newspapers in Finnish—and the Finnish written language—started maturing only in the latter half of the century. Consequently, we have been able to produce meaningful clusters of *isms* for 1820s onward for Swedish and only from the 1860s onward for Finnish.
- The *-ismi* was not a productive suffix in the Finnish language but used through cognate loans and through analogous derivation of foreign words. Consequently, *isms* are in general less common and *ism* words less productive in Finnish than in Swedish but nonetheless used especially as Finnish political language in the nineteenth century developed through an interplay between the two main languages in the country.
- The political outlook of the two languages was slightly different. From the 1880s onward the Finnish and Swedish newspapers were printed in nearly equal amounts. At this time the language spheres also started specializing. Swedish speakers lived mostly in larger towns and around the coast, whereas Finnish speakers occupied the whole country [16]. At this point, Finnish-language papers were more likely to have a rural or working-class background and Swedish-language papers were more likely to be more urban, liberal and bourgeois, which naturally also shows in the use of *isms*. This is typically visible in the proportionately big role the cluster around socialism manifests in Finnish compared to Swedish. The clusters clearly show how Finnish-language *ism* vocabulary was more politically oriented in the early twentieth century. Cultural, philosophical and scientific *isms* were less present. This has partly to do with the outlook of Finnish-language newspapers, but partly it seems that political *isms* were not as easily translated into vernacular forms without an *ism*, whereas for other terminology, this option was more readily at hand.

## 3.2 Politics and ideology as distinct clusters

Aligning the clusters in the Sankey plots provides a possibility of visually exploring how the vocabulary of *isms* developed over the course of the century. As can be seen in Figure 1, for Swedish there is quite a steady expansion of *isms* from the 1820s onward. As the models for producing the clusters rely on enough datapoints for training, particular clusters appear with a delay compared to first uses of particular words. For instance, patriotism appears the first time in the corpus in 1791 and liberalism 1820, but the clusters in which they are part of (but not necessarily cluster representatives or most frequent ones) appear in 1820–1839 and 1840–1859, as can be seen in Swedish clusters. The word socialism appears the first time in 1840 and also appears in the cluster for 1840–1859 respectively, since it immediately became popular and the amount of newspapers in Swedish had already grown.

Figure 1 suggests that there is a clear continuity in the politically laden *isms* which start from a cluster with *patriotism*, *fanatism* (Eng. fanaticism) and *despotism* in one cluster in 1820–1839 and continue with an expansion over the consecutive double decades. Most frequent *isms* in the political clusters are *patriotism*, *socialism* and *despotism* up to 1859, and then

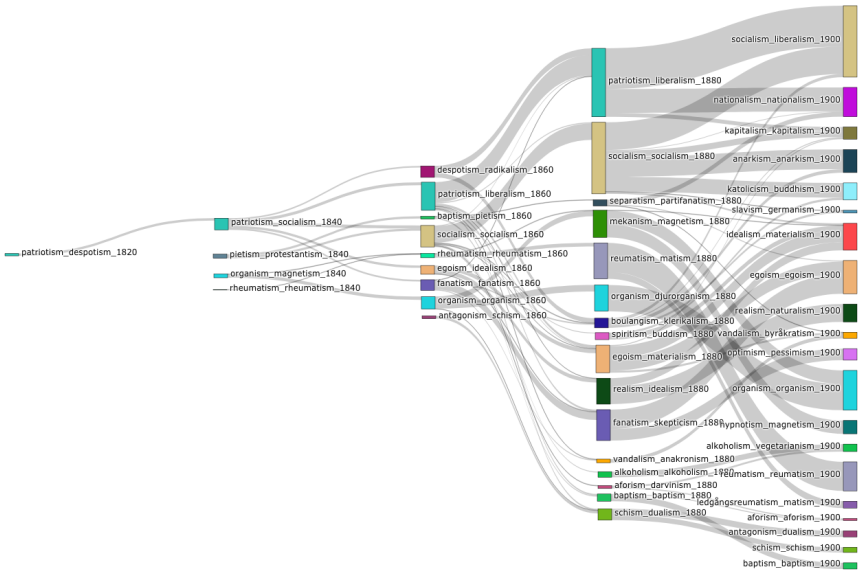
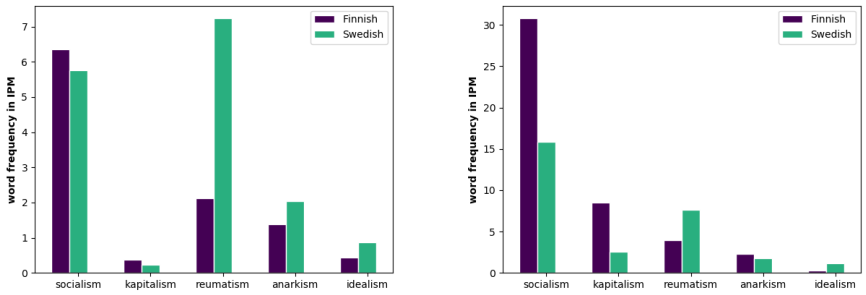



Figure 1: Sankey diagram of *ism* clusters from the Swedish dataset covering five double decades from 1820 to 1900. The cluster name is the most frequent ism word for that cluster followed by the cluster representative and the double decade.

*boulangism*, *fanatism*, *anarkism*, *nationalism* and *kapitalism* (Eng. capitalism) up to 1917. There is some fluctuation between the political clusters, like *liberalism* and *patriotism* being quite tightly associated until the last time slice of the investigated period, and some unsurprising continuities, like *konservatism* (Eng. conservatism) and *liberalism* being in the same clusters through out. Still, it seems that there is less fluctuation between the distinctly political clusters and the other clusters. Also the the religious *isms* (starting from *pietism*), and medical *isms* (*rheumatism*) come across as reasonably stable. The philosophical, artistic and scientific *isms* are also distinguishable, albeit they are less clear cut.

For Finnish, the data is too scarce to produce meaningful clusters for more than three time slices Even though the Finnish corpus for the 1880–1899 double decade is comparable in size with the Swedish corpus, the number of *distinct isms* in Finnish is smaller than in Swedish: 44 for Finnish and 125 for Swedish.

With scarcer data the distinctness of the clusters is even clearer. Clusters with socialism as the most frequent *ism* are rather dominant both for Swedish and Finnish, but the role of socialism as a pivotal *ism* is even more pronounced for the latter as is also indicated by [17]. Further work is needed to explain this in more detail, but apart from above mentioned demographic and political background factors for Finnish-language press, it also seems that the discourse on socialism may have been less confined in Finnish than in Swedish. Clustering the words with a cosine similarity to any *ism* word provides more information about the linguistic contexts of each *ism*. Table 2 shows how Finnish-language clusters with associated words includes more




(a) 1880–1889 (b) 1900–1917

Figure 2: Relative word frequencies (items per million) for selected isms. The labels are Swedish words (Finnish equivalents are *sosialismi*, *kapitalismi*, *reumatismi*, *anarkismi* and *idealismi*, respectively). For *reumatism* we sum up counts for two spelling variants (*reumatism* and *rheumatism*). Note that the plots have a different scale: counts in 1900–1917 are generally much larger.

religious (and to certain extent also scientific) terminology than the more political discourse visible in the Swedish-language clusters, the Finnish-language clusters include a more religious terminology for the period 1900–1917. Why socialist discourse was more prone to tap into a reservoir of religious rhetoric in Finnish than in Swedish requires further study.

Both Swedish-language and Finnish-language clusters include separate clusters for rheumatism (with spelling variations), which are almost self-containing. *Rheumatism*, albeit an *ism* based strictly on spelling, does not cluster with other *isms*, but has a distinct use in medical discourse of the time. This shows that our clustering method is effective, but it is also indicative of the fact that historical language use made a distinction of different types of *isms*. Some simply ended with the suffix, while others were seen as belonging to groups of other *isms*. *Rheumatism* also stands out as a specific type of term in the newspaper medium as it was very often used as a stand alone word in advertisements or lists of illnesses.

## 4 Discussion and Future Work

There are alternative ways to build diachronic embeddings. The recent line of research is aimed at smooth time representation [3, 7, 22, 26]. These methods reveal gradual semantic changes over the years instead of dividing the data into discrete time slices. In the future we plan to utilize one of these methods to investigate semantic drift of ideological terms in more details. We further aim to explore methods for cross-language cluster comparison. In the case of *ism* words, translations between Finnish and Swedish are near at hand as is clear in Figure 2a and 2b, but a proper comparison of the clusters needs further methodological exploration.

For examples see *Hufvudstadsbladet*, 23.11.1907, nro 320, p. 8; *Wiborgs Nyheter*, 23.01.1903, nro 18, p. 3; *Uusi Suometar*, 04.06.1905, nro 128, p. 8

Table 2: Enriched clusters for Finnish and Swedish that contain word *socialism*(*i*). Cluster *representatives* are marked with italic, **isms** are highlighted with bold.

| 1880–1889 | | | | 1900–1917 | | | |
|---|---|---|---|---|---|---|---|
| Finnish | | Swedish | | Finnish | | Swedish | |
| ***sosialismi*** | 5115 | ***socialism*** | 5560 | **sosialismi** | 75117 | ***socialism*** | 15080 |
| **anarkismi** | 1120 | reaktion | 6991 | kristitty | 72175 | socialdemokrati | 11030 |
| **nihilismi** | 602 | socialdemokrati | 2303 | *kristinusko* | 32542 | klasskamp | 2998 |
| **militarismi** | 328 | **anarkism** | 1975 | kristillisyys | 18566 | **anarkism** | 1709 |
| **kommunismi** | 316 | frigrelse | 1823 | rauhanaate | 1598 | socialdemokratien | 993 |
| **radikalismi** | 171 | proletariat | 1548 | **kommunismi** | 1548 | **absolutism** | 879 |
| sosiaalidemokratia | 386 | emancipation | 1225 | pakanakansa | 760 | framtidsstat | 533 |
| sosialidemokratia | 339 | **nihilism** | 1181 | buddhalaisuus | 456 | **individualism** | 512 |
| villitys | 337 | socialdemokratien | 1023 | lristinuslo | 428 | demokratien | 496 |
| luokkataistelu | 177 | utopi | 1016 | tinusko | 383 | **skandinavism** | 440 |
| reaktio | 136 | **antisemitism** | 911 | knnytys | 256 | **syndikalism** | 387 |
| pappis-malta | 130 | bourgeoisie | 772 | tristi | 252 | fredstank | 342 |
| | | anti | 747 | adventisti | 243 | **antisemitism** | 341 |
| | | elementerna | 703 | alliansi | 164 | **marxism** | 286 |
| | | **absolutism** | 641 | kristinuslo | 161 | **internationalism** | 285 |
| | | **klerikalism** | 569 | tinuslo | 147 | **antimilitarism** | 267 |
| | | **statssocialism** | 485 | buddalaisuu | 144 | **kommunism** | 256 |
| | | **kommunism** | 459 | tristinusko | 128 | historieuppfattning | 236 |
| | | **ateism** | 455 | jumalausko | 127 | studentrrelse | 170 |
| | | kvinnoemancipation | 445 | islami | 123 | **aktivism** | 168 |
| | | **panslavism** | 341 | buddalaisuusi | 119 | **revisionism** | 166 |
| | | reaktionen | 335 | konfusius | 118 | brandfackla | 142 |
| | | kvinnorrelse | 332 | lristinusko | 114 | kulturrrelse | 134 |
| | | framtidsstat | 242 | jrkeisoppi | 111 | frbudsrrelse | 122 |
| | | **kapitalism** | 226 | tristinuslo | 109 | **frsvarsnihilism** | 117 |
| | | **jesuitism** | 206 | alkukristillisyys | 103 | **nykterism** | 112 |
| | | **individualism** | 196 | | | **ungsocialism** | 112 |
| | | socia | 174 | | | **kollektivism** | 110 |
| | | ateistisk | 173 | | | **modernism** | 109 |
| | | fredsid | 155 | | | samhllsrrelse | 102 |
| | | **ultramontanism** | 129 | | | finskhetsrrelsen | 101 |
| | | **utilitarism** | 124 | | | | |
| | | kollektivistisk | 122 | | | | |
| | | **kollektivism** | 121 | | | | |
| | | **cesarism** | 110 | | | | |
| | | frihetsid | 108 | | | | |

## Acknowledgements

We are grateful to Simon Hengchen and Mark Granroth-Wilding for the help with data preparation. This work has been supported by the European Unions Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

## References

[1] Domagoj Alagić, Jan Šnajder, and Sebastian Padó. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] Cesare Cuttica. To use or not to use ... the intellectual historian and the isms: A survey and a proposal. *Études Épistémè*, 23, 2013.

[3] Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[4] Max Engman. *Språkfrågan: Finlandssvenskhetens uppkomst 1812-1922*. Svenska litteratursällskapet i Finland, 2016.

[5] Michael Freeden. *Ideology: A very short introduction*. Oxford University Press, 2003. OCLC: 312572349.

[6] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[7] Nabeel Gillani and Roger Levy. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *NAACL HLT 2019*, page 94, 2019.

[8] Simon Hengchen, Ruben Ros, and Jani Marjanen. A data-driven approach to the changing vocabulary of the nation in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In *In Proceedings of the Digital Humanities (DH) conference 2019, Utrecht, The Netherlands*, 2019.

[9] H. M. Höpfl. Isms. *British Journal of Political Science*, 13(1):1–17, 1983.

[10] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *ACL 2014*, page 61, 2014.

[11] Jussi Kurunmäki and Jani Marjanen. Isms, ideologies and setting the agenda for public debate. *Journal of Political Ideologies*, 23(3):256–282, 2018.

[12] Jussi Kurunmäki and Jani Marjanen. A rhetorical view of isms: an introduction. *Journal of Political Ideologies*, 23(3):241–255, 2018.

[13] Andrey Kutuzov, Elizaveta Kuzmenko, and Lidia Pivovarova. Clustering of Russian adjective-noun constructions using word embeddings. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 3–13, 2017.

[14] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, 2018.

[15] Eetu Mäkelä. Las: an integrated language analysis tool for multiple languages. *The Journal of Open Source Software*, 1, 2016.

[16] Jani Marjanen, Ville Vaara, Antti Kanner, Hege Roivainen, Eetu Mäkelä, Leo Lahti, and Mikko Tolonen. A national public sphere? analyzing the language, location, and form of newspapers in finland, 1771-1917. *Journal of European Periodical Studies*, 2019 (forthcoming).

[17] Wiktor Marzec and Risto Turunen. Socialisms in the Tsarist Borderlands. *Contributions to the History of Concepts*, 13(1):22–50, June 2018.

[18] Tomas Mikolov, Kai Chen, Greg S Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *NIPS*, 2013.

[19] Tuula Pääkkönen, Jukka Kervinen, Asko Nivala, Kimmo Kettunen, and Eetu Mäkelä. Exporting Finnish digitized historical newspaper contents for offline use. *D-Lib Magazine*, 22(7/8), 2016.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[22] Alex Rosenfeld and Katrin Erk. Deep neural models of semantic shift. In *NAACL HLT 2018*, pages 474–484, 2018.

[23] Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20:33–53, 2008.

[24] Ivo Spira. *A conceptual history of Chinese -isms: The modernization of ideological discourse, 1895-1925*. Number Volume 4 in Conceptual history and Chinese linguistics. Brill, 2015.

[25] Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*, 2018.

[26] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *The 11th ACM Conference on Web Search and Data Mining*, 2018.

# Appendix F: The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections

HAL
archives-ouvertes.fr

## The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections

Jani Marjanen, Jussi Kurunmäki, Lidia Pivovarova, Elaine Zosa

▶ **To cite this version:**

Jani Marjanen, Jussi Kurunmäki, Lidia Pivovarova, Elaine Zosa. The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections. 2020. hal-02491304

**HAL Id: hal-02491304**
**https://hal.inria.fr/hal-02491304**

Preprint submitted on 25 Feb 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal of Data Mining and Digital Humanities

episciences.org

# The expansion of isms, 1820–1917: Data-driven analysis of political language in digitized newspaper collections

**Jani Marjanen**[1], **Jussi Kurunmäki**[2], **Lidia Pivovarova**[1], **Elaine Zosa**[1]

[1]University of Helsinki, Finland
[2]Tampere University, Finland

Corresponding author: Jani Marjanen , jani.marjanen@helsinki.fi

**Abstract**

Words with the suffix -ism are reductionist terms that help us navigate complex social issues by using a simple one-word label for them. On the one hand they are often associated with political ideologies, but on the other they are present in many other domains of language, especially culture, science, and religion. This has not always been the case. This paper studies isms in a historical record of digitized newspapers from 1820 to 1917 published in Finland to find out how the language of isms developed historically. We use diachronic word embeddings and affinity propagation clustering to trace how new isms entered the lexicon and how they relate to one another over time. We are able to show how they became more common and entered more and more domains. Still, the uses of isms as traditions for political action and thinking stand out in our analysis.

**Keywords**

isms; ideology; political language; diachronic word embeddings; affinity propagation clustering

## I INTRODUCTION

Words with the suffix -ism are indispensable terms for understanding politics and society, yet they are complex words that give rise to plenty of confusions. It is hard to tell how different isms ranging from communism to Protestantism and further to impressionism and positivism really relate to one another. For sure, people using everyday language seem to uphold the link between isms, and from a analytical perspective it is clear that most words with the suffix serve some sort of reductionist function. They are words that describe something complex in just one heading [Spira, 2015].

Most studies on isms have tried to make sense of them by trying to create reasonable typologies for understanding their areas of application or characteristics [Cuttica, 2013, Höpfl, 1983]. This paper departs from a more historical take, by trying to understand the historical process in which they developed. Isms have been used to categorize things ever since antiquity. In English a separate word, isms, emerged in the seventeenth century to denote them collectively. Ever since the sixteenth and seventeenth centuries isms have spread to many new domains in life, covering everything from religion, politics, science, arts and more. Isms have also gained a global reach so that they are used as cognate loans or direct translations in many languages [Höpfl, 1983, Kurunmäki and Marjanen, 2018b, Spira, 2015].

By focusing on the nineteenth century and using digitized historical newspapers from Finland, this paper takes the historical view even further. It uses word embeddings to analyze the spread of isms in the Finnish context. This method drawn from natural language processing (NLP)

differs a lot from traditional approaches in history and political science, but the possibility of clustering isms in a relatively large historical data set has several benefits also for scholarship in the humanities and social sciences. It can partly confirm the narrative of isms becoming especially political and even ideological in the course of the nineteenth century, but also that isms relating to psychology and the sciences entered the lexicon at this time. The clustering clearly shows how these isms belonged to different language domains. Further, the method can point out interesting new findings about the scope and nature of particular isms and their use in the Finnish context. For instance, the role of the discourse on socialism is here charted in the context of semantically similar terms, as is the rhetoric of separatism in different domains of language.

The results regarding separatism also point toward potential new method development in using word embeddings to cluster semantically similar terms as the term had a clear semantic continuity but was readily used in different discourses in consecutive periods. This type of complex historical case present in diverse data such as newspapers may be a welcome challenge for testing and developing methods with contextualized word embeddings.

## II RESEARCH QUESTIONS AND DATA

### 2.1 Research questions

This paper studies isms as particularly laden keywords in societal discourse in Finland in the long nineteenth century. We address the following research questions:

- How did the vocabulary of isms expand in the period?
- Which isms appear as similar based on their embeddings?
- How does the theme of politics distinguish itself in the clusters of isms over time?
- Are there interesting continuities in the enriched clustering that takes into account nearest neighbors of the isms?

Finally, we shortly discuss the differences in Finnish-language and Swedish-language discourse in Finland when looked upon through isms. It is worth pointing out that the bilingual nature of public discourse in nineteenth-century Finland creates a fruitful comparative starting point for investigating isms elsewhere in the world since the two languages were in constant interaction, but also developed at different speeds [Marjanen et al., 2019, Engman, 2016]. Similarly the transnational developments in the language of isms are heavily intertwined and follow the same trends, but local circumstances always set the context in which these words were used as political, social, and cultural keywords.

### 2.2 Data

To answer these questions, we use a digitalized collection of nineteenth-century Finnish newspapers freely available from the National Library of Finland [Pääkkönen et al., 2016]. Though the archive contains newspapers starting from 1770s, the earlier time periods do not have enough data for the analysis we apply in this paper. Thus, we keep to the data from 1820 to 1917. From 1809 to 1917 Finland was a Grand Duchy in the Russian empire and in this period it gained many state institutions of its own [Jussila, 2004]. The process of new political vocabulary entering the lexicon was heavily intertwined with the development of Finland as a state and a nation [Hyvärinen et al., 2003].

The collection contains newspapers in the Russian, German, Swedish and Finnish languages, with the latter two as the main languages. In our analysis, these dominant languages are treated as two separate corpora even though contemporaries often relied on newspapers in both languages. The period has been described as an interaction between three languages in Finland,

Table 1: Corpus size by double decade.

| Time slice | Millions of words | |
|---|---|---|
| | FINNISH | SWEDISH |
| 1820–1839 | 1.3 | 25.5 |
| 1840–1859 | 10.3 | 77.9 |
| 1860–1879 | 90.6 | 326.7 |
| 1880–1899 | 805.3 | 966.9 |
| 1900–1917 | 2439.0 | 953.0 |
| **Total** | 3346.6 | 2355.2 |

Swedish being the main language for administration and learned life, Finnish being the primary language of the majority of the inhabitants in Finland and increasingly seen as the language of the future, and Russian as the language that most people in Finland did not read, but still loomed in the background as the main language of the Russian empire [Engman, 2016].

In this paper we use the Finnish and Swedish corpora, leaving the far more smaller data sets of Russian and German for the further research. The total amount of words in both corpora is presented in Table 1. Both corpora are lowercased and lemmatized using LAS, an open-source language-analysis tool [Mäkelä, 2016].[1] It is a meta-analysis tool that provides a wrapper for other existing tools developed for specific tasks and languages. Though LAS supports multiple languages, most efforts were done to process Finnish data, including historical Finnish. The output for our Swedish data is more noisy. In particular, the Swedish LAS lemmatizer is unable to predict the lemma for out-of-vocabulary words, e.g. *boulangismen* (definite form of 'boulangism'). Thus we applied additional normalization by converting all words ending with *-ismen* or *-ismens* into *-ism* forms. For all other words we use the LAS output; implementation of proper Swedish lemmatization is beyond the scope of this paper, as most of our findings are based on clustering the isms only, thus perfect lemmatization of other words is less crucial.

## III METHOD

### 3.1 Diachronic embeddings

To trace semantic shifts in word meanings we split a lemmatized corpus into double decades (1820–1839, 1840–1859, and so on until 1900–1917) and train continuous embeddings [Mikolov et al., 2013] on each time slice. We use the Gensim Word2Vec implementation [Řehůřek and Sojka, 2010] using the Skip-gram model, with a vector dimensionality of 100, window size 5 and a frequency threshold of 100—only lemmas that appear more than 100 times within a double decade are used for training. In this way we try to ensure that each word in a model has a reliable amount of context and the embeddings are trustworthy. However, we lose some isms because they appear less than 100 times in a double-decade. For example, the Finnish word *feminismi* was mentioned 91 times between 1900 and 1917 and was excluded from our analysis, while its Swedish counterpart was mentioned 242 times and is visible in our results. Our models allow us to detect when a word became frequent, in what context it was used and what is the difference between the Swedish and Finnish contexts. They do not allow us, however, to check when the word appeared for the first time and comparison of word distributions between languages is not fully reliable for less frequent words.

Since training word embeddings is a stochastic process, the particular values of vectors do

[1] https://github.com/jiemakel/las

not stay close across runs, though distances between words are quite stable. To ensure that embeddings are stable across time slices, we follow the vector initialization approach proposed in [Kim et al., 2014]: embeddings for $t+1$ time slice are initialized with vectors built on $t$; then training continues using new data. The learning rate value is set to the end learning rate of the previous model, to prevent models from diverging rapidly. This approach has been previously used in [Hengchen et al., 2019] with slightly different data.

Temporally aligned embeddings have been used before to trace semantic drift by computing distances between vectors representing a word in two time periods or by measuring differences in nearest neighbours for these vectors [Hamilton et al., 2016]. However, most studies that tackle semantic shift detection in computational linguistics deals with clear cases of word meaning change such as the complete change of meaning of the word 'gay' or acquiring of a new completely different sense such as words 'virus' or 'cell'. These rapid transformations could also be found in our data: e.g. Swedish word *flygare*, which initially meant an insect but changed its meaning to "aviator" in the beginning of the twentieth century. The embedding models that we trained is able to detect this change, since the nearest neighbors of *flygare* completely changed. At the same time, distance-based methods seem to be less useful for isms, since their meanings do not change to that extreme. For example, 'patriotism', whether it had positive or negative connotations, always has a meaning semantically close to "love of one's country". At the same time, the political and social context in which the word was used changed over time. Further the term could be used for quite different rhetorical purposes and it carried new social and affective meanings that are not as readily visible in the embeddings.[2] Thus, in this paper we do not lean on distances between word vectors across time and instead use clustering to find which isms were closer to each other—i. e., had similar contexts—in various periods of time.

### 3.2 Clustering

To investigate the expansion of the vocabulary of isms we cluster words into close groups based on their embeddings. Since our task is mostly exploratory and the number of clusters cannot be known in advance we apply the Affinity Propagation clustering technique [Frey and Dueck, 2007]. The method splits all datapoints into *exemplars*, i.e., cluster representative tokens, and *instances*, i.e., other members of clusters. At the initial step all datapoints present a cluster of their own. Then for each instance-representative pair a likelihood for an instance to be represented by an exemplar is computed by taking into account all other instances of the exemplar and all other available exemplars for the instance. This computation is repeated until convergence is reached; if an exemplar has no instances it is dismissed. We use the standard implementation of this algorithm from the Scikit-learn package [Pedregosa et al., 2011] with default parameters.

Affinity Propagation has been previously used for various language analysis tasks, including collocation clustering into semantically related classes [Kutuzov et al., 2017] and unsupervised word sense induction [Alagić et al., 2018]. The main advantages of the method are that it detects the number of clusters automatically and is able to produce clusters of various size. As a side effect it returns exemplars, i.e. cluster representatives, which are not necessarily equal to the geometric centre of the cluster.

The main drawback of the Affinity Propagation is pairwise computations. The method is quadratic in time and memory and cannot be applied to large datasets, such as a whole corpus vocabulary. Thus, data selection is an unavoidable step. In this paper we use Affinity Propagation in two experiments.

[2] For social, affective and other types of meaning, see Leech [1974]

In the first experiment, we extract from the corpus all ism words. i.e. words that end with *-ism* in Swedish and *-ismi* in Finnish and cluster only this set of words. We exclude from the list words that are shorter than 5 characters for Swedish and 6 characters for Finnish. This is to filter out obvious errors that appeared in the optical character recognition of the newspapers such as 'ism', 'tism', or 'rism'. Though the words 'ism' and 'ismi' exist in the Swedish and Finnish languages, they are very uncommon in nineteenth-century press. The extraction allows us to identify how close these words are to each other given other isms in the corpus.

In the second experiment, we try to put isms into a richer context and trace other words associated with them in the respective double-decades. We extract from the corpus all words which have a cosine similarity to any isms that is less than 0.5. Then we perform clustering on this enriched dataset. Finally, the clusters are filtered so that only clusters that contain at least one isms word are presented for qualitative analysis. An output of this procedure is different compared to the first experiment, i.e. words that were clustered together in the isms-only clustering, can break up into different enriched clusters, since in the latter setting they have more exemplar options.

Henceforth we refer to the results of the first and the second experiments as *ism clusters* and *enriched clusters* respectively. We discuss the outcomes of the two experiments interchangingly since they give different perspectives on the development of ism vocabulary.

Clustering is performed separately for each time slice. To link clusters across time we perform visualization with Sankey charts. In the Sankey diagram, clusters from time slice $t$ are linked to clusters in time slice $t + 1$ if they have words in common. The magnitude of the link is the sum of the word frequencies of the common words between the linked clusters from adjacent time slices. We use the frequencies from the source cluster, that is the cluster from time slice $t$.

## IV RESULTS

Some of our results are directly related to the political history of Finland and the development of newspapers as a medium, whereas others go well together with previous notions of the development of the language of isms in general. They strengthen earlier interpretations by giving more robust proof for interpretations that have mostly relied on the qualitative reading of sources. Other findings come across as surprising also for historians of political ideologies, and may at least to some extent force us to rethink how we look upon the history of political discourse. In what follows, we will present the findings in this order.

### 4.1 Swedish and Finnish clusters in comparison

As expected, Finnish-language and Swedish-language isms cluster differently in terms of timing and themes that are present (see Figure 1 and Figure 2). There are three main reasons for this:

- Swedish-language press in Finland developed earlier and included more abstract content earlier in the century, whereas newspapers in Finnish—and the Finnish written language—started maturing only in the latter half of the century. Consequently, we have been able to produce meaningful clusters of isms for 1820s onward for Swedish and only from the 1860s onward for Finnish. As described earlier, the languages were in constant interaction, but the scope of Finnish-language newspapers was much smaller in the first half of the century and the content was to a lesser degree theoretical and political. Furthermore, Swedish-language newspapers were quicker in adopting new terms from publications in Sweden because of the language connection and thus had a sort of mediating function with regard to new political vocabulary.
- The *-ismi* was not a productive suffix in the Finnish language but used through cognate

FINNISH

| Time slice | *ism* | *close* | *cluster* | *select* |
|---|---|---|---|---|
| 1820 - 1839 | 0 | - | - | - |
| 1840 - 1859 | 0 | - | - | - |
| 1860 - 1879 | 1 | 157 | 1 | 12 |
| 1880 - 1899 | 35 | 5977 | 20 | 442 |
| 1900 - 1917 | 119 | 8940 | 70 | 1543 |

SWEDISH

| Time slice | *ism* | *close* | *cluster* | *select* |
|---|---|---|---|---|
| 1820 - 1839 | 3 | 724 | 3 | 49 |
| 1840 - 1859 | 17 | 1845 | 12 | 211 |
| 1860 - 1879 | 61 | 5229 | 31 | 669 |
| 1880 - 1899 | 120 | 12233 | 54 | 1320 |
| 1900 - 1917 | 137 | 11858 | 56 | 1387 |

Table 2: Number of distinct words used on various steps of the to obtain enriched clusters: **isms** is a number of distinct words with suffix *-ism*, ***close*** is a number of words, which cosine similarity to at least one ism is higher than 0.5, ***cluster*** is a number of clusters that contain at least one ism, ***select*** is a number of words in these clusters.

loans and through analogous derivation of foreign words.[3] Consequently, isms are in general less common in Finnish than in Swedish. Nonetheless they were used in both languages especially as Finnish political language developed through an interplay with Swedish. In the particular case of adopting isms as key terminology in Finnish, the latter half of the century was a crucial turning point.

- The political outlook of the two languages was slightly different. From the 1880s onward the Finnish and Swedish newspapers were printed in nearly equal amounts. At this time the language spheres also started specializing. Swedish speakers lived mostly in larger towns and around the coast, whereas Finnish speakers inhabited most of the country [Marjanen et al., 2019]. In Lapland, Sami languages also had a strong presence, but they were not at this time published in print. At this point, Finnish-language papers were more likely to have a rural or working-class background and Swedish-language papers were more likely to be more urban, liberal and bourgeois, which also shows in the use of isms. This is typically visible in the proportionately big role the cluster around socialism manifests in Finnish compared to Swedish. The clusters clearly show that Finnish-language ism vocabulary was more politically oriented in the early twentieth century. Cultural, philosophical and scientific isms were less present.

The distinction between Swedish and Finnish is also visible from the analysis of the enriched clusters. The number of words used on various steps of analysis is presented in Table 2, which shows that the number isms in the Finnish data is much smaller than for the Swedish data. The table also shows that though 0.5 is an arbitrary threshold, up to 90% of words selected using this threshold are filtered out after the clustering. This is an indirect justification that the threshold is sufficient and most of the relevant words are present in the output. The number of selected clusters is generally smaller than the number of words with the suffix ism since they tend to cluster together.

[3] As such the ism is not strictly speaking a suffix in Finnish, but a rather a sublexical suffix-like unit as often the whole words a cognate loans in which the root itself is not a word in Finnish. We thank Antti Kanner for pointing this out.

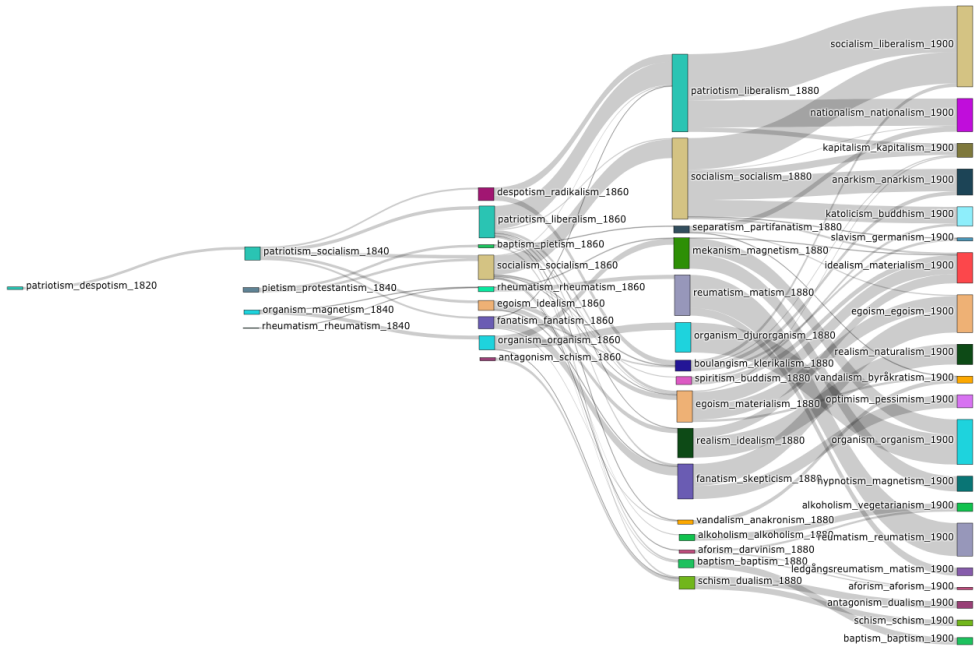



Figure 1: Sankey diagram of isms clusters from the Swedish dataset covering five double decades from 1820 to 1917. The cluster name is the most frequent ism word for that cluster followed by the cluster representative and the double decade.

### 4.2 Expansion of the language of isms

By looking at the raw counts of different isms we see an expansion of isms in the nineteenth century. This is partly the function of a growth in data size over time, but mostly because new isms were introduced and often also lexicalized to the extent that they became nodal points in newspaper discourse. One feature of the suffix is that it is rather easy to deploy in *ad hoc* inventions of new words, so many isms were introduced, but never resonated in public use. These are as such interesting instances of linguistic innovation, but are excluded in this study as we use a frequency threshold for training our embeddings. The threshold also effectively excludes many false variants caused by noisy optical character recognition.

Aligning the clusters in the Sankey plots provides a possibility of visually exploring how the vocabulary of isms developed over the course of the century. As can be seen in Figure 1, there is quite a steady expansion of isms from the 1820s onward for Swedish. As the models for producing the clusters rely on enough datapoints for training, particular clusters appear with a delay compared to first uses of particular words. For instance, patriotism appears the first time in the corpus in 1791 and liberalism 1820, but the clusters in which they are part of (but not necessarily cluster representatives or most frequent ones) appear in 1820–1839 and 1840–1859, as can be seen in Swedish clusters (Table 8). The word socialism appears the first time in 1840 and is also included in the cluster for 1840–1859, since it immediately became popular and the amount of newspapers in Swedish had already grown.

The visualization of Finnish-language clusters provides a much shorter story, but the expansion

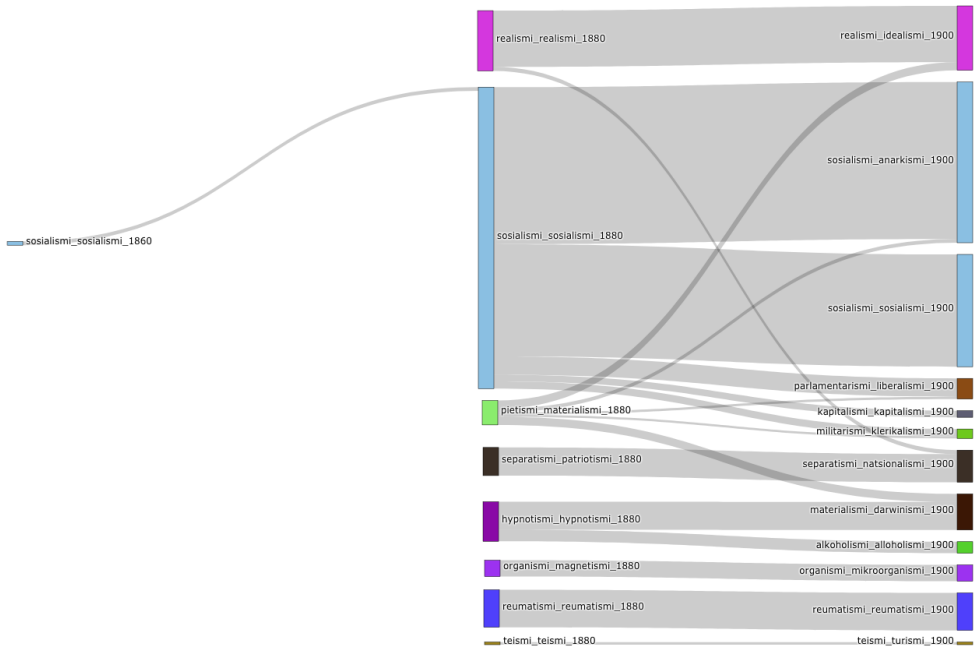



Figure 2: Sankey diagram of isms clusters from the Finnish dataset covering three double decades from 1860 to 1917. The cluster name is the most frequent ism word for that cluster followed by the cluster representative in the double decade.

of isms into new domains is also visible in this data. A peculiarity of the Finnish data is that although the number of isms grew also for Finnish, the clusters show much stronger continuities in the sense that the clusters fluctuate much less than for the Swedish data. In this case the emergence of clusters that can be described as political or ideological are the ones that change the landscape of isms most significantly, whereas medical, cultural and scholarly isms only play a minor role.

### 4.3 Politics and ideology as distinct clusters

Previous interpretations by Kurunmäki and Marjanen [2018a], have suggested that the early nineteenth century meant the breakthrough of isms that we today associate with major political ideologies, whereas the end of the century saw the rise of plenty of new isms in the sciences (including medicine) and the arts. Again, looking at first appearances of particular isms in the Swedish-language data set suggests this holds also for Finland. However, the clusters allow for a stronger claim that suggest that the political and ideological isms form a quite distinct category after they have been introduced.

Figure 1 and Table 8 show that there is a clear continuity in the politically laden isms which start from a cluster with *patriotism*, *fanatism* (Eng. fanaticism) and *despotism* in one cluster in 1820–1839 and continue with an expansion over the consecutive double decades. Most frequent isms in the political clusters are *patriotism*, *socialism* and *despotism* up to 1859, and then *boulangism*, *fanatism*, *anarkism*, *nationalism* and *kapitalism* (Eng. capitalism) up to 1917. There is some fluctuation between the political clusters, like *liberalism* and *patriotism* being quite tightly associated with one another until the last time slice of the investigated period, and

some unsurprising continuities, like *konservatism* (Eng. conservatism) and *liberalism* being in the same clusters throughout. Still, it seems that there is less fluctuation between the distinctly political clusters and the other clusters. Also religious isms (starting from *pietism*), and medical isms (e.g. *rheumatism*) come across as fairly stable. The philosophical, artistic and scientific isms are also distinguishable, albeit they do cluster more freely. The case of rheumatism is very specific as it has a high frequency and appears often in health-related advertisements, which means it does not co-occur very often with other isms, but is rather an isolated marketing term in marketing pills and ointments.

For Finnish-language, the data is too scarce to produce meaningful clusters for more than three time slices Even though the Finnish corpus for the 1880–1899 double decade is comparable in size with the Swedish corpus, the number of *distinct isms* in Finnish is smaller than in Swedish: 44 for Finnish and 125 for Swedish.

With scarcer data the distinctness of the clusters is even clearer. Clusters with socialism as the most frequent ism are rather dominant both for Swedish and Finnish, but the role of socialism as a pivotal ism is even more pronounced for the latter as is also indicated by Marzec and Turunen [2018]. Further work is needed to explain this in more detail, but apart from above mentioned demographic and political background factors for Finnish-language press, it also seems that the discourse on socialism may have been less confined in Finnish than in Swedish.

### 4.4 Socialism as a pivotal ism

While the two data sets are different, they both show that a many isms pivot around the discourse of socialism especially toward the end of the century. Socialism does not fluctuate between clusters, but really seems to be one of the terms that organized the debate. We get supplementary perspective on this phenomenon by looking at the relative frequency of a selection of most frequent isms in our data (Figure 3). Like the clusters, the relative frequencies indicate a growing proportion of isms over time and also demonstrate some difference between the data sets. For the Swedish data set we see a change on the overall landscape of the vocabulary with terms such as patriotism being dominant at first but then surpassed in frequency by socialism. In Swedish, we also find a broader selection of isms from political to religious and medical topics, present for the second half of the nineteenth century. In Finnish, the landscape is different as it appears that the whole vocabulary relating to isms was dominated by socialism from the 1860s onward. It appears as if the word socialism in a way invited other isms to be lexicalized in the Finnish language. Once socialism became inevitable in Finnish-language political discourse, other isms well-known from Swedish and other Germanic languages were easier to introduce also to Finnish. This does not mean that isms did not at all feature in Finnish, only that they were infrequent and not a normal part of the lexicon. We must also note that most authors who produced texts in Finnish, also operated in Swedish, so while they did not write about isms in Finnish, they still held notions of isms through the other main language of the country.

Albeit a comparison with China may come across as far fetched, the introduction of isms in central categories for political thinking in China provides a dramatic instance that can be contrasted to the Finnish case. As Ivo Spira has shown, the discussion on modernization in China in the early 1900s prompted comparisons with Western ideological discourse through a discussion of different isms. At this time, a sign corresponding to ism was introduced (*zhǔyì* 主義) and it quickly became a way of translating Western isms into Chinese, but also a way to conceptualize locally embedded isms [Spira, 2018]. As such, the isms may be seen as a way of synchronizing Chinese and Western political thought, so that they were used to translate and compare ideological positions [Jordheim, 2014, 2017].

Table 3: Enriched clusters for Finnish and Swedish that contain word *socialism*(*i*). Cluster *representatives* are marked with italic, **isms** are highlighted with bold.

| 1880–1889 | | | |
|---|---|---|---|
| FINNISH | | SWEDISH | |
| **sosialismi 'socialism'** | 5115 | **socialism 'socialism'** | 5560 |
| ***anarkismi* 'anarchism'** | 1120 | *reaktion* 'reaction' | 6991 |
| ***nihilismi* 'nihilism'** | 602 | *socialdemokrati* 'social democracy' | 2303 |
| ***militarismi* 'militarism'** | 328 | ***anarkism* 'anarchism'** | 1975 |
| ***kommunismi* 'communism'** | 316 | *frigörelse* 'liberation' | 1823 |
| ***radikalismi* 'radicalism'** | 171 | *proletariat* 'proletariat' | 1548 |
| *sosiaalidemokratia* 'social democracy' | 386 | *emancipation* 'emancipation' | 1225 |
| *sosialidemokratia* 'social democracy' | 339 | ***nihilism* 'nihilism'** | 1181 |
| *villitys* 'craze' | 337 | *socialdemokratien* 'social democracy' | 1023 |
| *luokkataistelu* 'class struggle' | 177 | *utopi* 'utopia' | 1016 |
| *reaktio* 'reaction' | 136 | ***antisemitism* 'antisemitism'** | 911 |
| *pappis-malta* 'clericalism'$_{ocr}$ | 130 | *bourgeoisie* 'bourgeoisie' | 772 |
| | | *anti* 'anti-'$_{ocr}$ | 747 |
| | | *elementerna* 'elements' | 703 |
| | | ***absolutism* 'absolutism'** | 641 |
| | | ***klerikalism* 'clericalism'** | 569 |
| | | ***statssocialism* 'state socialism'** | 485 |
| | | ***kommunism* 'communism'** | 459 |
| | | ***ateism* 'atheism'** | 455 |
| | | *kvinnoemancipation* 'women's emancipation' | 445 |
| | | ***panslavism* 'panslavsim'** | 341 |
| | | *reaktionen* 'reaction' | 335 |
| | | *kvinnorörelse* 'women's movement' | 332 |
| | | *framtidsstat* 'future state' | 242 |
| | | ***kapitalism* 'capitalism'** | 226 |
| | | ***jesuitism* 'jesuitism'** | 206 |
| | | ***individualism* 'individualism'** | 196 |
| | | *socia* 'social'$_{ocr}$ | 174 |
| | | *ateistisk* 'atheistic' | 173 |
| | | *fredsidé* 'idea of peace'$_{ocr}$ | 155 |
| | | ***ultramontanism* 'ultramontanism'** | 129 |
| | | ***utilitarism* 'utilitarianism'** | 124 |
| | | *kollektivistisk* 'collectivistic' | 122 |
| | | ***kollektivism* 'collectivism'** | 121 |
| | | ***cesarism* 'cesarism'** | 110 |
| | | *frihetsidé* 'idea of liberty' | 108 |

Turning back to the issue of socialism as a pivotal ism in both Swedish- and Finnish-language discourse in Finland, our findings harmonize with Marzec and Turunen [2018] who emphasize the role of socialism based on frequency and textual analysis, but we further note that looking at socialism in the context of all isms shows that it also had a synchronizing function between Finnish and Swedish. The breakthrough of socialism as a buzz word in the second half of the nineteenth century helped produce political and ideological isms also in Finnish that could be compared with counterparts in Swedish and other languages.

Careful analysis of text would provide more reliable interpretations to why socialism gained such a dominant role in Finnish-language discourse, but our enriched clustering with a cosine similarity to any word does also provide more information about the linguistic contexts of each

Table 4: Enriched clusters for Finnish and Swedish that contain word *socialism*(*i*). Continuation.

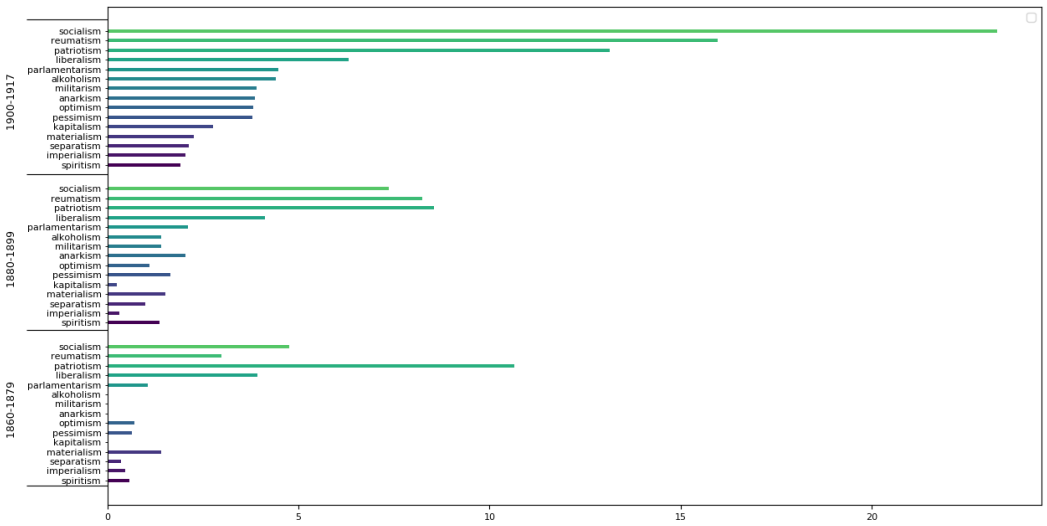
| 1900–1917 | | | |
|---|---|---|---|
| FINNISH | | SWEDISH | |
| ***sosialismi* 'socialism'** | 75117 | **socialism 'socialism'** | 15080 |
| *kristitty* 'christian' | 72175 | *socialdemokrati* 'social democracy' | 11030 |
| kristinusko 'christianity' | 32542 | *klasskamp* 'class struggle' | 2998 |
| *kristillisyys* 'christian' | 18566 | ***anarkism* 'anarchism'** | 1709 |
| *rauhanaate* 'pacifism' | 1598 | *socialdemokratien* 'social democracy' | 993 |
| ***kommunismi* 'communism'** | 1548 | ***absolutism* 'absolutism'** | 879 |
| *pakanakansa* 'pagan people' | 760 | *framtidsstat* 'future state' | 533 |
| *buddhalaisuus* 'buddhism' | 456 | ***individualism* 'individualism'** | 512 |
| *lristinuslo* 'christianity'$_{ocr}$ | 428 | *demokratien* 'democracy' | 496 |
| *tinusko* 'christianity'$_{ocr}$ | 383 | ***skandinavism* 'skandinavism'** | 440 |
| *käännytys* 'conversion' | 256 | ***syndikalism* 'syndicalism'** | 387 |
| *tristi* '?'$_{ocr}$ | 252 | *fredstank* 'pacifism' | 342 |
| *adventisti* 'adventist' | 243 | ***antisemitism* 'antisemitism'** | 341 |
| *alliansi* 'alliance' | 164 | ***marxism* 'marxism'** | 286 |
| *kristinuslo* 'christianity'$_{ocr}$ | 161 | ***internationalism* 'internationalism'** | 285 |
| *tinuslo* 'christianity'$_{ocr}$ | 147 | ***antimilitarism* 'antimilitarism'** | 267 |
| *buddalaisuu* 'buddhism'$_{ocr}$ | 144 | ***kommunism* 'communism'** | 256 |
| *tristinusko* 'christianity'$_{ocr}$ | 128 | *historieuppfattning* 'understanding of history' | 236 |
| *jumalausko* 'faith' | 127 | *studentrörelse* 'student movement' | 170 |
| *islami* 'islam' | 123 | ***aktivism* 'activism'** | 168 |
| *buddalaisuusi* 'buddhism'$_{ocr}$ | 119 | ***revisionism* 'revisionism'** | 166 |
| *konfusius* 'confucius' | 118 | *brandfackla* 'bombshell' | 142 |
| *lristinusko* 'christianity'$_{ocr}$ | 114 | *kulturrörelse* 'cultural movement' | 134 |
| *järkeisoppi* 'philosophy' | 111 | *förbudsrörelse* 'prohibition movement' | 122 |
| *tristinuslo* 'christianity' | 109 | ***försvarsnihilism* 'defence nihilism'** | 117 |
| *alkukristillisyys* 'early christianity' | 103 | ***nykterism* 'prohibition movement'** | 112 |
| ***ungsocialism* 'ungsocialism'** | 112 | | |
| ***kollektivism* 'collectivism'** | 110 | | |
| ***modernism* 'modernism'** | 109 | | |
| *samhällsrörelse* 'social movement' | 102 | | |
| *finskhetsrörelsen* 'finnish movement' | 101 | | |

ism. Tables 3 and 4 show how Finnish-language clusters with words associated with socialism include more religious (and to certain extent also scientific) terminology than the more political discourse visible in the Swedish-language clusters. Why socialist discourse was more prone to tap into a reservoir of religious rhetoric in Finnish than in Swedish requires further study. One possible explanation to this may lie in the fact that socialism was in Finnish to a higher degree than Swedish related more strongly to the so-called social question, that is social an political problematization of class issues, poverty and labor issues and that these issues also dovetailed with Finnish-language religious discourse around the turn of the century 1900.

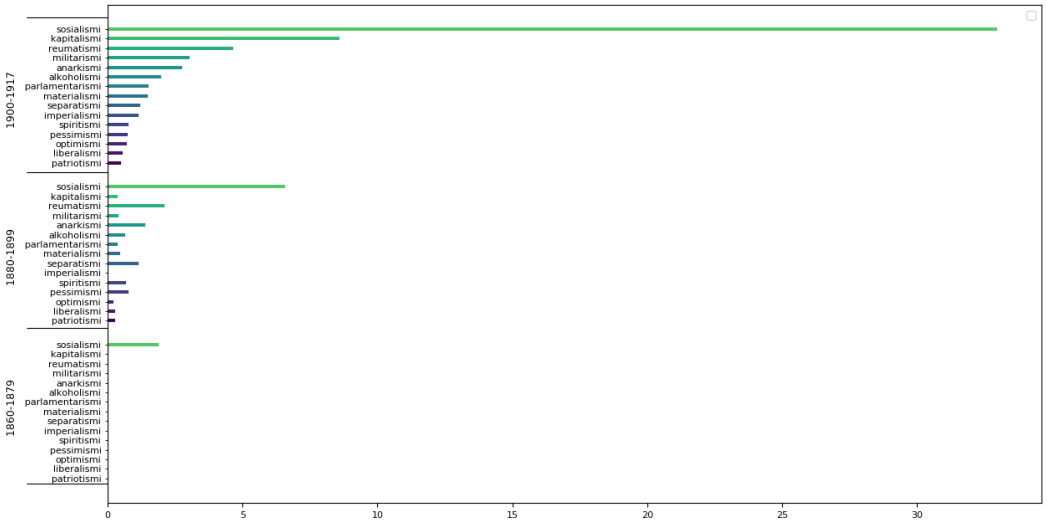
### 4.5 Separatism and its different domains

If words like socialism and rheumatism shows remarkable continuity through clusters, other isms seem to be less tied to their clusters. A surprising and illuminating example of this is separatism in both Swedish and Finnish. In Table 5, we present the enhanced clusters for it in the Swedish data set.

Most of the words similar to separatism in the 1860–1879 cluster are religious, philosophical or scientific notions, such as mysticism, Darwinism, human nature, negation or idealistic. By




SWEDISH




FINNISH

Figure 3: A selection of the most frequent words ending with suffix *-ism/ismi*. The x-axis presents relative frequency in items per million.

analyzing the clusters and reading sample texts from the period, we conclude that the cluster derives much from debates about religion and the historical experience of Lutheranism being threatened. In the period new scientific and philosophical strands of thought as well as contemporary religious revival movements seriously challenged the status of the dominant state church in Finland, and the notion of separatism seems to have been readily used in the ensuing debates.

The 1880–1899 cluster contains completely different set of words, including reference to ethnicity and language policy in the country, such as Finnishness, Fennomans and language policy, and contains rather emotional expressions, such as *agitation* and *fanaticism*. The outlier of pho-

tophobia also belongs to a similar discourse as the term was used metaphorically at the time to discuss things that could not be brought to the fore because of political tensions. Again with selected reading of texts we note that separatism is clearly clustered with words that are related to a contemporary discussion about national identity and national language.

The 1900–1917 cluster is again different from the previous two and contains more general political lexis. Again, it seems that the notion of separatism had been included in a new discoursive domain. Now, the word clusters with words that relate to state structures and even the context of the Russian empire. Separatism had become embedded in discussions about independence, the role of Finland and as a nation.

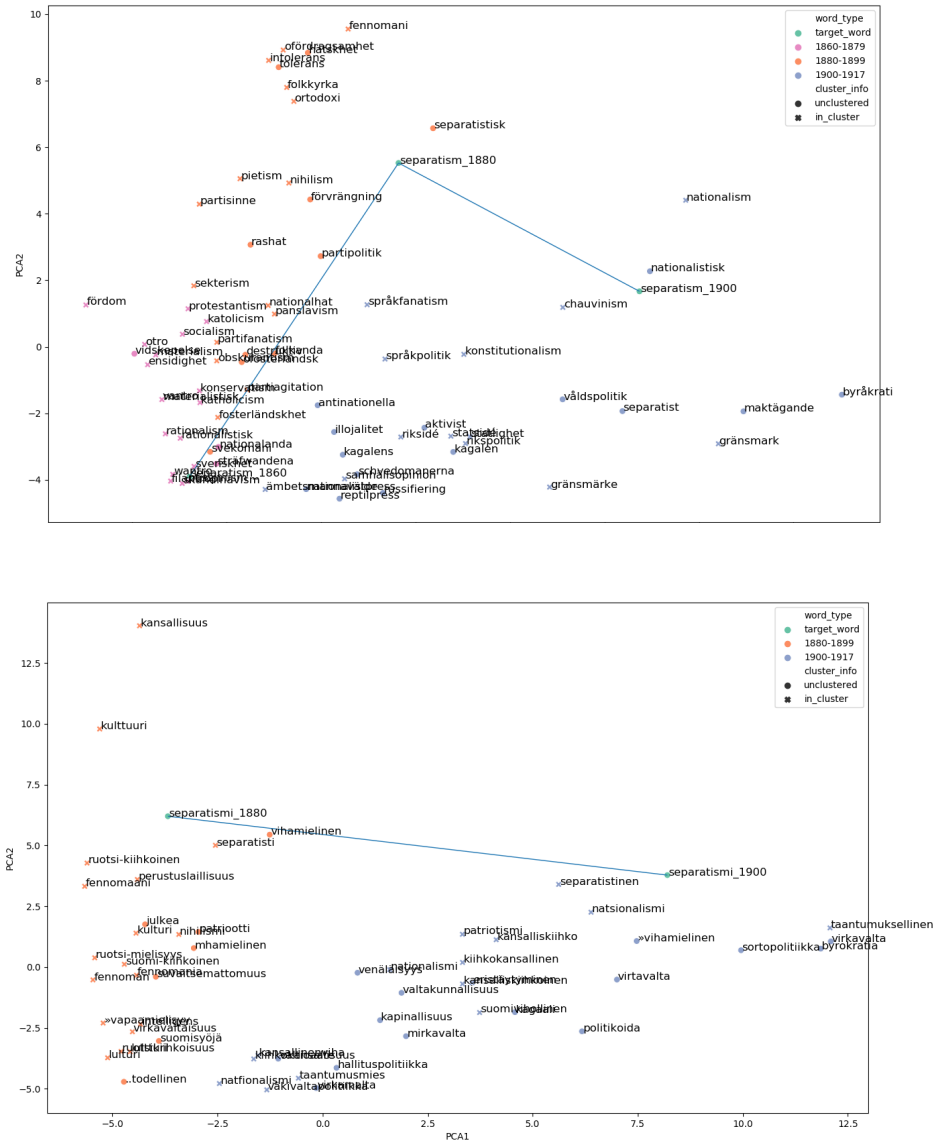
All in all, in three consecutive double decades separatism at first had a mostly religious context, when it was adopted into a discourse relating to ethnicity and the language question which is so central to the period, and finally it spread into a more general political discourse in which Finlad's status as a part of the Russian empire was discussed.

The Finnish-language clusters for *separatismi*, presented in Table 6, suggest a similar development, but given the language divide in the country, the perspective is slightly different. The Finnish data set does not include a cluster for the period 1860–1879 as the word occurs less than a hundred times and as a consequence is excluded from our models. The periods for 1880–1899 and 1900–1917 point at separatism that is first dominated by the language question and then in the early twentieth century being dominated by the issue of Finland's status in the empire and nationalism in general. Interestingly, however, the Finnish-language cluster for 1880–1899 contains more words that relates to the Svekomans, that is the Swedish-language movement, and the Swedish-language cluster includes more words relating to Fennomans, that is the Finnish-language movement. Together with a reading of a selection of the sources, we see how the discourse on separatism is quite similar in both languages, but is directed toward the "opposing" side. (Obviously, the language question was interwoven with issues of class and the urban-rural divide as well, so it is not as simple as talking about only two sides, but the general pattern is clear.) The Finnish-language cluster for the period 1900–1917 is also clearly similar to the Swedish-language counterpart, but again the vocabulary presents two different, but overlapping, perspectives.

The change in the distribution of *separatism* seems to be related to a change in the dominant context in which it was discussed (from a religious context to a political context). The shift in cluster entails some degree of semantic change, but it is also clear that separatism as a highly abstract term could lend itself to many different themes or topics, and thus it seems the change in dominant themes themselves is more important for the changing clusters than the changes in meaning of the word. An alternative interpretation would be that separatism was a polysemous word in which the different separatisms (those relating to religion, the language issue or the national question) coincided and that different senses dominated in different time slices, but a reading of sample sentences does not support this interpretation.

The distributional shift of separatism is to some extent visible from changes in the nearest neighbours of the word presented in Figure 4. They visualize a shift from the time slice 1880–1899 to 1900–1917 in both languages. The outlook can be interpreted in a similar way as the clusters produced by Affinity Propagation, but have a slightly different selection of words.

This can be explained by the nature of the procedure used to produce the visualization. PCA is a dimensionality reduction technique and does not explicitly do any clustering therefore each word can be among the nearest neighbours for any number of other words while Affinity Prop-




FINNISH

Figure 4: PCA plots of *separatism(i)* and its nearest neighbours across time slices. Words marked by × are part of the separatism cluster in their respective time slice.

agation assigns a word to exactly one cluster so that, for instance, *socialism* and *katolicism* are separated in clusters of their own. The difference between outputs demonstrates an added value of the clustering, which selects only one word split among many possibilities provided by embeddings. At the same time, this also means loss of information, especially for polysemous words.

| 1860-1879 | 1880-1899 | 1900-1917 |
|---|---|---|
| *separatism* ‘ separatism’<br>*mysticism* ‘ mysticism’ *naturalism* ‘ naturalism’<br>*darwinism* ‘ darwinism’ *moral* ‘ morality’<br>*tidsanda* ‘ zeitgeist’ *krass* ‘ crass’ *utopi* ‘ utopia’<br>*materialistisk* ‘ materialistic’ *otro* ‘ incredible’<br>*rationalistisk* ‘ rationalistic’ *wantro* ‘misbelief’<br>*menniskonaturen* ‘ human nature’ *tidehvarfvets* ‘the age (genitive)’<br>*materialism* ‘ materialism’ *materialist* ‘ materialistic’<br>*konservatism* ‘ conservatism’<br>*idealism* ‘ idealism’ *rationalism* ‘ rationalism’<br>*negation* ‘ negation’ *abstraktion* ‘ abstraction’<br>*idealistisk* ‘idealistic’ | *separatism* ‘ separatism’ *rent* ‘?’<br>*finskhet* ‘ Finnishness’ *fennomanins* ‘ Fennomania’<br>*fennomani* ‘ Fennomania’ *svenskhet* ‘ Swedishness’<br>*fennomanin* ‘ Fennomania’ *vikingaparti* ‘ Viking party’<br>*språkpolitik* ‘ language policy’ *publicistisk* ‘ journalistic’<br>*partiagitation* ‘ party agitation’ *partiyra* ‘party delirium’<br>*partifanatism* ‘ party fanaticism’<br>*språkgräl* ‘ language quarrel’<br>*språkfanatism* ‘ language fanaticism’<br>*språkfråga* ‘ language question’<br>*språkfrägan* ‘language question’<br>*ljusskygghet* ‘ photophobia’ | *separatism* ‘ separatism’ *riksidé* ‘ national idea’$_{ocr}$<br>*statsidé* ‘ state idea’$_{ocr}$ *rikspolitik* ‘ national policy’<br>*bourgeoisins* ‘ bourgeoisie’ *byråkratien* ‘ bureaucracy’<br>*samhällsopinion* ‘ societal opinion’<br>*sträfvandenas* ‘aspirations’ *rikskomplex* ‘ national complex’<br>*nationalitet-* ‘ nationality’$_{ocr}$ *santryska* ‘ true Russian’$_{ocr}$<br>*ämbetsmannavälde* ‘ officialdom’<br>*gränsmärke* ‘ borderline’ *gränsmark* ‘ borderline’$_{ocr}$<br>*riksenhet* ‘ national assembly’<br>*samhällskraft* ‘ social force’ *statlighet* ‘ statehood’<br>*frihetssträvande* ‘freedom-aspiring’ *wäldets* ‘domination/empire’<br>*riksmakt* ‘ national power’ *självhärskarmakten* ‘autocratic power’ |

Table 5: Swedish clusters containing word *separatism*

| 1880-1899 | 1900-1917 |
|---|---|
| *separatismi* ‘separatism’ *ruotsi-kiihkoinen* ‘Svekoman’ *ruotsinmielinen* ‘Swedish-minded’<br>*ruotsalaisuus* ‘Swedishness’ *viikinki* ‘Viking’ *ruotsi-mielinen* ‘Swedish-minded’<br>*fennomaani* ‘Fennoman’ *epäkansallinen* ‘anti-national’ *viikingit* ‘Vikings’<br>*separatisti* ‘separatist’ *ruotsikko* ‘Swedish-minded’(person) *miikinki* ‘Viking’$_{ocr}$ *pöppö* ‘?’<br>*miikingit* ‘Vikings’$_{ocr}$ *suomimielinen* ‘Finnish-minded’ *ruotsi-mielisyys* ‘Swedish-mindedness’<br>*wiitinki* ‘Viking’$_{ocr}$ *wiilinki* ‘Viking’$_{ocr}$ *miitinki* ‘Viking’$_{ocr}$ *ruotsimielinen* ‘Swedish-minded’<br>*suomi-kiihkoinen* ‘Fennoman’ *fennoman* ‘Fennoman’ *henkiheimolainen* ‘like minded’<br>*dagbladilainen* ‘member of the Dagblad circle’ *miiking* ‘Viking’$_{ocr}$ *fennomani* ‘Fennoman’<br>*wiiking* ‘Viking’$_{ocr}$ *fennomaaninen* ‘Fennoman’ *ruotsikiihkoisuus* ‘Svekomania’<br>*wiilinli* ‘Viking’$_{ocr}$ *miikinkilehti* ‘Vikings’ newspaper’$_{ocr}$ *suomenmielinen* ‘Finnish-minded’$_{ocr}$<br>*miikinkiläinen* ‘Viking (adjective)’$_{ocr}$ *ruolsinmielinen* ‘Swedish-minded’ *ruotsiliihloinen* ‘Svekoman’$_{ocr}$<br>*herranenluokka* ‘class of the lords’ *miikingilehti* ‘Vikings’ newspaper’$_{ocr}$ *epälansallinen* ‘anti-national’$_{ocr}$ | *separatismi* ‘separatism’<br>*nationalismi* ‘nationalism’ *natsionalismi* ‘nationalism’<br>*opportunismi* ‘opportunism’ *natfionalismi* ‘nationalism’$_{ocr}$<br>*eristäytyminen* ‘isolation’ *kansalliskiihko* ‘nationalism’<br>*intelligens* ‘intelligence’ *länsieurooppalainen* ‘Western-European’<br>*rotutaistelu* ‘race struggle’ *vapaamielisyy* ‘liberalism’$_{ocr}$<br>*sanomalehdistö!* ‘press’ *antipatia* ‘antipathy’<br>*kansallinenviha* ‘national anger’ *kiihkokansallisuus* ‘national fervour’<br>*eristäytyä* ‘self-isolate’ *liittolaisuus* ‘alliance’<br>*vihamieli-syy* ‘hostility’$_{ocr}$ *kansallinenylpeys* ‘national pride’<br>*kielipolitiikka* ‘language policy’<br>*kansallinenliike* ‘national movement’ |

Table 6: Finnish clusters containing word *separatismi*

# V DISCUSSION AND FUTURE WORK

## 5.1 Embeddings and semantics

As we have shown in this paper, the comparison of word embeddings trained on various time periods is a fruitful method for analysis of historical newspapers. Diachronic analysis using vector models is a rapidly growing research field in computational linguistics (see, for example, recent surveys of this topic [Kutuzov et al., 2018, Tahmasebi et al., 2018]).

The most recent research involves using contextual word embeddings reviewed in Ethayarajh [2019] and exemplified in BERT [Devlin et al., 2019] and ELMo [Peters et al., 2018]. They output a separate vector for each word mentioned based on its context. These models make possible tracing differences in word usage across time, though as far as we are aware these models were applied to trace an evolution of a single word—e.g., [Martinc et al., 2020a,b]—rather than detecting evolution of groups of semantically related words.

Another research direction is aimed at continuous time representation [Dubossarsky et al., 2019, Gillani and Levy, 2019, Rosenfeld and Erk, 2018, Yao et al., 2018]. These methods reveal gradual semantic changes over time and do not require dividing the data into discrete time slices.

Finally, much effort is contributed into development of cross-lingual embeddings [Ruder et al., 2019], which put words from two or more languages into the same vector space and thus enable direct comparison of data from various languages. We suggest that using any of these approaches—namely, contextual, continuous and cross-lingual embeddings—or their combination might be a productive next step, which would allow us to deeper understand historical development of complex political notions.

## 5.2 Digital humanities and the study of political vocabularies

The analysis of the history of political thought is not tied to the newest advances in natural language processing, but analyses drawing on them often create space for new interpretations in studying the political imaginaries of past people. In this study on isms as nodes of everyday political thinking in nineteenth-century newspapers from Finland, we have produced new and

reliable ways of charting and visualizing the expansion of the vocabulary of isms. Especially noteworthy in our method is that it can grasp developments in word use that relate both to growth in frequency and change in the distribution of the word. This way our findings regarding the importance of socialism as a political keyword are not surprising to someone with good knowledge of the political vocabulary in Finland, but our method shows the sheer amounts and pivotal role of socialism in a way that has not been possible before. Nor has there been any attempts to compare the discourse of socialism across the language divide in Finland. The findings relating to separatism are different in the sense that we were not expecting to find anything out of ordinary relating to that word. We were rather surprised that it emerged as a interesting case based on a semi data-driven perspective.

Our cases relating to socialism and separatism also indicate that the relationship between distribution and meaning as pointed out in the so-called distributional hypothesis [Sahlgren, 2008] is not as straightforward as sometimes believed.[4] While there is a link between the change in distribution and semantic change, this link seems to be easier to capture in clear cases of polysemy than in rather vague and flexible terms such as the isms under study here. Isms are often also in hierarchical relations to one another, especially when being qualified in some way. for instance, the words state socialism (*statssocialism*) and municipal socialism (*kommunalsocialism*) found in 8 of which the former clusters together with socialism, but not the latter, suggests that the clustering is rather being related to social meaning than to strict conceptual meaning.

While word embeddings and other methods analysing the distribution of terminology are increasingly looking for new avenues in studying multilingual corpora, we further want to point out that the case of isms may be a fruitful avenue for developing multilingual approaches. Dealing with Finnish and Swedish in one country showed that the historical translatability between the language (even if Finnish is less prone to introduce new isms) can be very useful in studying political vocabularies and thinking in different linguistic contexts — when combined with good contextual knowledge that takes into account linguistic an political specificites relating to the languages at stake.

## VI ACKNOWLEDGEMENTS

We are grateful to Simon Hengchen and Mark Granroth-Wilding for the help with data preparation. This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDDIA).

## References

Domagoj Alagić, Jan Šnajder, and Sebastian Padó. Leveraging lexical substitutes for unsupervised word sense induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Cesare Cuttica. To use or not to use ... the intellectual historian and the isms: A survey and a proposal. *Études Épistémè*, 23, 2013. ISSN 1634-0450. doi: 10.4000/episteme.268. URL http://journals.openedition.org/episteme/268.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, 2019.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Max Engman. *Språkfrågan: Finlandssvenskhetens uppkomst 1812-1922*. Svenska litteratursällskapet i Finland, 2016. ISBN 978-951-583-354-9.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT,

[4] We thank Antti Kanner for pointing this interpretation out to us.

ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://www.aclweb.org/anthology/D19-1006.

Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814): 972–976, 2007.

Nabeel Gillani and Roger Levy. Simple dynamic word embeddings for mapping perceptions in the public sphere. In *NAACL HLT 2019*, page 94, 2019.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access, 2016.

Simon Hengchen, Ruben Ros, and Jani Marjanen. A data-driven approach to the changing vocabulary of the 'nation' in English, Dutch, Swedish and Finnish newspapers, 1750-1950. In *In Proceedings of the Digital Humanities (DH) conference 2019, Utrecht, The Netherlands*, 2019.

H. M. Höpfl. Isms. *British Journal of Political Science*, 13(1):1–17, 1983. ISSN 1469-2112, 0007-1234. doi: 10.1017/S0007123400003112. URL https://www.cambridge.org/core/journals/british-journal-of-political-science/article/isms/36C46DBC3F69FB510E33A0D6C6888DFF.

Matti Hyvärinen, Jussi Kurunmäki, Kari Palonen, Tuija Pulkkinen, and Henrik Stenius, editors. *Käsitteet liikkeessä: Suomen poliittisen kulttuurin käsitehistoria*. Vastapaino, Tampere, 2003. ISBN 978-951-768-130-8.

Helge Jordheim. Introduction: Multiple times and the work of synchronization. *History and Theory*, 53(4):498–518, 2014.

Helge Jordheim. Synchronizing the world: Synchronism as historiographical practice, then and now. *History of the Present*, 7(1):59–95, 2017.

Osmo Jussila. *Suomen suuriruhtinaskunta 1809–1917*. WSOY, Helsinki, 2004. ISBN 978-951-0-29500-7.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *ACL 2014*, page 61, 2014.

Jussi Kurunmäki and Jani Marjanen. Isms, ideologies and setting the agenda for public debate. *Journal of Political Ideologies*, 23(3):256–282, 2018a. doi: 10.1080/13569317.2018.1502941. URL https://www.tandfonline.com/doi/full/10.1080/13569317.2018.1502941.

Jussi Kurunmäki and Jani Marjanen. A rhetorical view of isms: an introduction. *Journal of Political Ideologies*, 23(3):241–255, 2018b. ISSN 1356-9317, 1469-9613. doi: 10.1080/13569317.2018.1502939. URL https://www.tandfonline.com/doi/full/10.1080/13569317.2018.1502939.

Andrey Kutuzov, Elizaveta Kuzmenko, and Lidia Pivovarova. Clustering of Russian adjective-noun constructions using word embeddings. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 3–13, 2017.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, 2018.

Geoffrey N. Leech. *Semantics*. Penguin, Harmondsworth, 1974. ISBN 978-0-14-021694-3.

Eetu Mäkelä. Las: an integrated language analysis tool for multiple languages. *The Journal of Open Source Software*, 1, 2016.

Jani Marjanen, Ville Vaara, Antti Kanner, Hege Roivainen, Eetu Mäkelä, Leo Lahti, and Mikko Tolonen. A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917. *Journal of European Periodical Studies*, 4(1):54–77, June 2019. ISSN 2506-6587. doi: 10.21825/jeps.v4i1.10483. URL https://neruda0.ugent.be/jeps/article/view/10483.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan*, 2020a.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. In *LREC*, 2020b.

Wiktor Marzec and Risto Turunen. Socialisms in the Tsarist Borderlands. *Contributions to the History of Concepts*, 13(1):22–50, June 2018. ISSN 1807-9326, 1874-656X. doi: 10.3167/choc.2018.130103. URL http://berghahnjournals.com/view/journals/contributions/13/1/choc130103.xml.

Tomas Mikolov, Kai Chen, Greg S Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *NIPS*, 2013.

Tuula Pääkkönen, Jukka Kervinen, Asko Nivala, Kimmo Kettunen, and Eetu Mäkelä. Exporting Finnish digitized

historical newspaper contents for offline use. *D-Lib Magazine*, 22(7/8), 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

Alex Rosenfeld and Katrin Erk. Deep neural models of semantic shift. In *NAACL HLT 2018*, pages 474–484, 2018.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631, 2019.

Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20:33–53, 2008.

Ivo Spira. *A conceptual history of Chinese -isms: The modernization of ideological discourse, 1895-1925*. Number Volume 4 in Conceptual history and Chinese linguistics. Brill, 2015. ISBN 978-90-04-28787-7.

Ivo Spira. Chinese isms: the modernization of ideological discourse in China. *Journal of Political Ideologies*, 23(3):283–298, September 2018. ISSN 1356-9317, 1469-9613. doi: 10.1080/13569317.2018.1502937. URL https://www.tandfonline.com/doi/full/10.1080/13569317.2018.1502937.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*, 2018.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *The 11th ACM Conference on Web Search and Data Mining*, 2018.

## A ANNEX 1: ISM CLUSTERS FOR FINNISH DATA

Table 7: Clustering obtained for Finnish words ending with ism suffix. We show cluster words and their frequencies in the respective time slice, sorted by frequency. Cluster *representatives* are marked with italic.

**1860-1879**

| | |
|---|---|
| *sosialismi* | 172 |

**1880-1899**

| | |
|---|---|
| *sosialismi* | 5115 |
| anarkismi | 1120 |
| nihilismi | 602 |
| militarismi | 328 |
| kommunismi | 316 |
| parlamentarismi | 312 |
| kapitalismi | 301 |
| liberalismi | 236 |
| radikalismi | 171 |
| boulangismi | 163 |
| kosmopolitismi | 128 |
| protestantismi | 115 |
| baptismi | 108 |

| | |
|---|---|
| *realismi* | 1029 |
| pessimismi | 614 |
| idealismi | 351 |
| symbolismi | 291 |
| naturalismi | 279 |
| optimismi | 182 |

| | |
|---|---|
| separatismi | 924 |
| *patriotismi* | 231 |
| fanatismi | 126 |

| | |
|---|---|
| organismi | 415 |
| *magnetismi* | 328 |

| | |
|---|---|
| pietismi | 370 |
| *materialismi* | 367 |
| ateismi | 167 |
| metodismi | 147 |
| dualismi | 105 |
| despotismi | 101 |

| | |
|---|---|
| *hypnotismi* | 733 |
| spiritismi | 542 |
| alkoholismi | 527 |

| | |
|---|---|
| *reumatismi* | 1706 |

| | |
|---|---|
| *teismi* | 119 |

**1900-1917**

| | |
|---|---|
| realismi | 2097 |
| pessimismi | 1192 |
| *idealismi* | 591 |
| naturalismi | 794 |
| pietismi | 476 |
| impressionismi | 342 |
| aforismi | 262 |
| humanismi | 242 |
| symbolismi | 231 |
| panteismi | 230 |
| egoismi | 201 |
| kubismi | 169 |
| asketismi | 154 |
| fatalismi | 152 |
| altruismi | 150 |
| mystisismi | 141 |
| klassisismi | 123 |
| ratsionalismi | 119 |

| | |
|---|---|
| materialismi | 3232 |
| spiritismi | 1327 |
| hypnotismi | 623 |
| monismi | 449 |
| *darwinismi* | 435 |
| modernismi | 174 |
| marxismi | 148 |
| pragmatismi | 113 |

| | |
|---|---|
| organismi | 1009 |
| magnetismi | 433 |
| *mikro-organismi* | 130 |

| | |
|---|---|
| *mekanismi* | 564 |

| | |
|---|---|
| sosialismi | 75117 |
| *anarkismi* | 5630 |
| terrorismi | 2063 |
| kommunismi | 1548 |
| sialismi | 910 |
| syndikalismi | 524 |
| individualismi | 398 |
| nihilismi | 397 |
| ateismi | 341 |
| antimilitarismi | 306 |
| revisionismi | 288 |
| sofalismi | 178 |
| remisionismi | 139 |
| indimidualismi | 127 |
| rialismi | 117 |
| darvinismi | 111 |
| antisemitismi | 102 |

| | |
|---|---|
| *>sosialismi* | 832 |
| ..sosialismi | 316 |
| sionismi | 221 |
| vegetarismi | 196 |
| ”sosialismi | 170 |
| .sosialismi | 163 |
| sosialismi | 133 |
| ’sosialismi | 126 |
| sosialismi | 108 |

| | |
|---|---|
| *reumatismi* | 9629 |
| matismi | 180 |
| nivelreumatismi | 158 |

| | |
|---|---|
| militarismi | 7062 |
| imperialismi | 2796 |
| despotismi | 661 |
| absolutismi | 350 |
| tsarismi | 318 |
| huliganismi | 270 |
| panslavismi | 202 |
| vandalismi | 194 |
| bolshevismi | 194 |
| hellenismi | 187 |
| *klerikalismi* | 147 |
| klerkalismi | 146 |
| germanismi | 104 |

| | |
|---|---|
| separatismi | 2008 |
| *natsionalismi* | 1852 |
| optimismi | 1580 |
| patriotismi | 993 |
| nationalismi | 657 |
| fanatismi | 589 |
| nalismi | 134 |
| anakronismi | 129 |
| natfionalismi | 120 |

| | |
|---|---|
| fotsialismi | 203 |
| *fofalismi* | 151 |
| anarfismi | 126 |
| fofialismi | 123 |

| | |
|---|---|
| *onnismi* | 517 |
| onanismi | 265 |

| | |
|---|---|
| *kapitalismi* | 20681 |
| talismi | 388 |
| industrialismi | 346 |
| suurkapitalismi | 302 |
| barbarismi | 225 |
| tpitalismi | 178 |
| lpitalismi | 166 |
| tapitalismi | 155 |
| pitalismi | 137 |
| lapitalismi | 113 |
| kapitalismi | 104 |

| | |
|---|---|
| parlamentarismi | 3413 |
| *liberalismi* | 1156 |
| radikalismi | 645 |
| feodalismi | 438 |
| opportunismi | 420 |
| dualismi | 293 |
| valtiososialismi | 235 |
| protestantismi | 213 |
| gmerkantilismi | 140 |

| | |
|---|---|
| alkoholismi | 4312 |
| kunnallinensosialismi | 1085 |
| *alloholismi* | 105 |
| holismi | 158 |

| | |
|---|---|
| teismi | 598 |
| tarismi | 175 |
| *turismi* | 126 |

| | |
|---|---|
| *reumatismi* | 212 |
| .reumatismi | 122 |

## B ANNEX 2

Table 8: Clustering obtained for Swedish words ending with ism suffix. We show cluster words and their frequencies, sorted by frequency. Cluster *representatives* are marked with italic.

**1820-1839**

| | |
|---|---|
| patriotism | 232 |
| fanatism | 165 |
| *despotism* | 153 |

**1840-1859**

| | |
|---|---|
| patriotism | 581 |
| egoism | 560 |
| fanatism | 431 |
| *socialism* | 294 |
| pauperism | 290 |
| despotism | 254 |
| kommunism | 160 |
| liberalism | 136 |
| radikalism | 105 |

| | |
|---|---|
| organism | 410 |
| mekanism | 217 |
| *magnetism* | 170 |
| galvanism | 124 |

| | |
|---|---|
| rheumatism | 101 |

| | |
|---|---|
| pietism | 505 |
| *protestantism* | 342 |
| katholicism | 155 |

**1860-1879**

| | |
|---|---|
| patriotism | 2664 |
| *liberalism* | 1148 |
| materialism | 461 |
| konservatism | 446 |
| dualism | 347 |
| parlamentarism | 341 |
| absolutism | 265 |
| optimism | 228 |
| pessimism | 209 |
| rationalism | 158 |
| konstitutionalism | 139 |
| anakronism | 137 |
| protektionism | 135 |
| sofism | 100 |

| | |
|---|---|
| *rheumatism* | 571 |
| reumatism | 305 |
| galvanism | 151 |

| | |
|---|---|
| *socialism* | 1263 |
| katolicism | 988 |
| protestantism | 846 |
| kommunism | 363 |
| nihilism | 282 |
| katholicism | 272 |
| mormonism | 240 |
| pauperism | 211 |
| skandinavism | 211 |
| jesuitism | 207 |
| spiritism | 188 |
| panslavism | 163 |
| darwinism | 112 |
| ateism | 102 |

| | |
|---|---|
| baptism | 360 |
| *pietism* | 205 |
| separatism | 118 |

| | |
|---|---|
| despotism | 923 |
| *radikalism* | 585 |
| ultramontanism | 458 |
| bonapartism | 337 |
| klerikalism | 177 |
| imperialism | 150 |
| carlism | 141 |
| federalism | 136 |
| republikanism | 115 |

| | |
|---|---|
| *fanatism* | 1710 |
| terrorism | 325 |
| vandalism | 233 |
| cynism | 190 |
| chauvinism | 115 |

| | |
|---|---|
| antagonism | 336 |
| *schism* | 323 |

| | |
|---|---|
| egoism | 1024 |
| realism | 388 |
| *idealism* | 183 |
| heroism | 153 |
| mysticism | 142 |
| naturalism | 115 |
| dilettantism | 106 |

| | |
|---|---|
| *organism* | 1575 |
| mekanism | 992 |
| magnetism | 328 |
| statsorganism | 120 |

Table 8: Clustering obtained for Swedish words ending with ism suffix: continuation

**1880-1899**

| | |
|---|---|
| *socialism* | 5560 |
| katolicism | 2154 |
| anarkism | 1975 |
| protestantism | 1408 |
| militarism | 1366 |
| nihilism | 1181 |
| antisemitism | 911 |
| absolutism | 641 |
| statssocialism | 485 |
| kommunism | 459 |
| journalism | 244 |
| bimetallism | 212 |
| jesuitism | 206 |
| nationalism | 198 |
| individualism | 196 |
| utilitarism | 124 |
| germanism | 115 |

| | |
|---|---|
| *baptism* | 641 |
| mormonism | 503 |
| sekterism | 366 |
| metodism | 259 |
| finlandism | 223 |
| laestadianism | 132 |
| fennicism | 106 |

| | |
|---|---|
| separatism | 829 |
| *partifanatism* | 278 |
| språkfanatism | 273 |
| nepotism | 121 |

| | |
|---|---|
| vandalism | 572 |
| *anakronism* | 298 |

| | |
|---|---|
| patriotism | 4792 |
| *liberalism* | 2705 |
| konservatism | 1806 |
| parlamentarism | 1688 |
| radikalism | 1455 |
| protektionism | 1222 |
| chauvinism | 950 |
| despotism | 868 |
| opportunism | 344 |
| skandinavism | 311 |
| konstitutionalism | 259 |
| republikanism | 203 |
| feodalism | 101 |

| | |
|---|---|
| mekanism | 3237 |
| hypnotism | 1811 |
| *magnetism* | 932 |
| idiotism | 287 |
| jordmagnetism | 175 |
| galvanism | 169 |
| somnambulism | 138 |
| bypnotism | 132 |
| atavism | 106 |

| | |
|---|---|
| schism | 1263 |
| antagonism | 863 |
| *dualism* | 467 |
| statsorganism | 146 |

| | |
|---|---|
| aforism | 283 |
| darwinism | 276 |
| *darvinism* | 165 |
| vegetarianism | 154 |

| | |
|---|---|
| egoism | 3057 |
| *materialism* | 1003 |
| pietism | 547 |
| formalism | 482 |
| ateism | 455 |
| rationalism | 276 |
| obskurantism | 221 |
| positivism | 221 |
| indifferentism | 213 |
| industrialism | 136 |
| asketism | 131 |
| barbarism | 123 |

| | |
|---|---|
| realism | 2295 |
| naturalism | 1134 |
| *idealism* | 834 |
| symbolism | 561 |
| mysticism | 422 |
| dilettantism | 309 |
| sofism | 309 |
| humanism | 216 |
| kosmopolitism | 171 |

| | |
|---|---|
| spiritism | 1123 |
| kannibalism | 383 |
| *buddism* | 175 |
| muhamedanism | 166 |
| buddhaism | 151 |
| spiritualism | 103 |

| | |
|---|---|
| *alkoholism* | 1364 |
| pauperism | 231 |
| morfinism | 129 |

| | |
|---|---|
| boulangism | 1128 |
| terrorism | 707 |
| *klerikalism* | 569 |
| panslavism | 341 |
| kapitalism | 226 |
| hellenism | 206 |
| partikularism | 180 |
| imperialism | 151 |
| bonapartism | 143 |
| ultramontanism | 129 |
| kollektivism | 121 |
| cesarism | 110 |

| | |
|---|---|
| fanatism | 3086 |
| pessimism | 1382 |
| cynism | 846 |
| optimism | 839 |
| *skepticism* | 548 |
| heroism | 320 |
| fatalism | 310 |
| lokalpatriotism | 112 |

| | |
|---|---|
| reumatism | 5735 |
| ledgångsreumatism | 1381 |
| rheumatism | 1262 |
| *matism* | 274 |
| ledgångsrheumatism | 188 |
| ledgängsrheumatism | 126 |

| | |
|---|---|
| organism | 5713 |
| mikroorganism | 621 |
| *djurorganism* | 110 |

| | |
|---|---|
| *amerikanism* | 161 |

Table 8: Clustering obtained for Swedish words ending with ism suffix: continuation

**1900-1917**

| | |
|---|---|
| socialism | 15080 |
| parlamentarism | 2231 |
| *liberalism* | 2034 |
| konservatism | 2034 |
| imperialism | 1637 |
| radikalism | 1438 |
| absolutism | 879 |
| klerikalism | 818 |
| konstitutionalism | 695 |
| protektionism | 650 |
| skandinavism | 440 |
| opportunism | 288 |
| marxism | 286 |
| internationalism | 285 |
| proportionalism | 245 |
| demokratism | 234 |
| statssocialism | 204 |
| aktivism | 168 |
| oktobrism | 136 |
| försvarsnihilism | 117 |
| monarkism | 112 |
| modernism | 109 |

| | |
|---|---|
| alkoholism | 2829 |
| vegetarism | 245 |
| darwinism | 193 |
| kommunalsocialism | 145 |
| *vegetarianism* | 130 |
| nykterism | 112 |

| | |
|---|---|
| antagonism | 943 |
| *dualism* | 415 |
| statsorganism | 177 |
| parallellism | 105 |

| | |
|---|---|
| *schism* | 3237 |
| skism | 167 |

| | |
|---|---|
| *turism* | 448 |

| | |
|---|---|
| idealism | 1113 |
| *materialism* | 694 |
| individualism | 512 |
| spiritism | 506 |
| pietism | 415 |
| mysticism | 266 |
| sofism | 260 |
| journalism | 189 |
| humanism | 179 |
| indifferentism | 178 |
| kosmopolitism | 167 |
| rationalism | 158 |
| obskurantism | 156 |
| ateism | 146 |
| asketism | 138 |
| atavism | 129 |
| dogmatism | 121 |
| monism | 114 |

| | |
|---|---|
| realism | 1785 |
| *naturalism* | 560 |
| impressionism | 319 |
| symbolism | 247 |
| dilettantism | 245 |
| kubism | 225 |
| klassicism | 183 |

| | |
|---|---|
| slavism | 317 |
| *germanism* | 240 |
| panslavism | 211 |
| hellenism | 141 |
| pangermanism | 130 |

| | |
|---|---|
| *reumatism* | 6423 |
| rheumatism | 820 |
| muskelreumatism | 142 |

| | |
|---|---|
| ledgångsreumatism | 1811 |
| *matism* | 470 |

| | |
|---|---|
| *anarkism* | 1709 |
| terrorism | 1600 |
| syndikalism | 387 |
| antisemitism | 341 |
| antimilitarism | 267 |
| kommunism | 256 |
| feminism | 242 |
| jesuitism | 180 |
| revisionism | 166 |
| nihilism | 125 |
| ungsocialism | 112 |
| kollektivism | 110 |

| | |
|---|---|
| *nationalism* | 5398 |
| patriotism | 4254 |
| separatism | 1079 |
| chauvinism | 1002 |
| språkfanatism | 373 |
| suometarianism | 363 |
| fariseism | 134 |

| | |
|---|---|
| katolicism | 1363 |
| protestantism | 726 |
| kannibalism | 338 |
| buddism | 261 |
| *buddhism* | 154 |
| muhammedanism | 150 |

| | |
|---|---|
| *organism* | 6627 |
| mekanism | 2332 |
| mikroorganism | 453 |
| samhällsorganism | 125 |

| | |
|---|---|
| optimism | 2565 |
| *pessimism* | 2023 |
| skepticism | 563 |

| | |
|---|---|
| *aforism* | 396 |
| finlandism | 221 |

| | |
|---|---|
| *egoism* | 2942 |
| fanatism | 2496 |
| cynism | 900 |
| heroism | 482 |
| fatalism | 252 |
| partifanatism | 219 |
| lokalpatriotism | 172 |
| altruism | 145 |
| klassegoism | 106 |
| knutpatriotism | 102 |

| | |
|---|---|
| *kapitalism* | 2399 |
| militarism | 2346 |
| despotism | 917 |
| industrialism | 732 |
| tsarism | 568 |
| barbarism | 169 |
| utilitarism | 142 |
| feodalism | 132 |
| storkapitalism | 128 |

| | |
|---|---|
| vandalism | 703 |
| *byråkratism* | 426 |
| formalism | 496 |
| anakronism | 423 |
| nepotism | 117 |
| servilism | 106 |

| | |
|---|---|
| hypnotism | 528 |
| *magnetism* | 362 |
| idiotism | 133 |
| jordmagnetism | 116 |

| | |
|---|---|
| baptism | 207 |
| mormonism | 125 |
| sekterism | 121 |

| | |
|---|---|
| *polism* | 182 |

# Appendix G: Dataset for Temporal Analysis of English-French Cognates

*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 855–859
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

## Dataset for Temporal Analysis of English-French Cognates

**Esteban Frossard[♠] Mickaël Coustaty[♠] Antoine Doucet[♠] Adam Jatowt[♢] Simon Hengchen[♣]**

[♠]University of La Rochelle, L3i Laboratory, {firstname.lastname}@univ-lr.fr
[♢]Kyoto University, adam@dl.kuis.kyoto-u.ac.jp
[♣]University of Helsinki, simon.hengchen@helsinki.fi

**Abstract**

Languages change over time and, thanks to the abundance of digital corpora, their evolutionary analysis using computational techniques has recently gained much research attention. In this paper, we focus on creating a dataset to support investigating the similarity in evolution between different languages. We look in particular into the similarities and differences between the use of corresponding words across time in English and French, two languages from different linguistic families yet with shared syntax and close contact. For this we select a set of cognates in both languages and study their frequency changes and correlations over time. We propose a new dataset for computational approaches of synchronized diachronic investigation of language pairs, and subsequently show novel findings stemming from the cognate-focused diachronic comparison of the two chosen languages. To the best of our knowledge, the present study is the first in the literature to use computational approaches and large data to make a cross-language diachronic analysis.

**Keywords:** Crosslingual semantic change, cognates, temporal analysis, semantic analysis

### 1. Introduction

Languages, our main tools of communication, evolve constantly: words obtain new and lose old meanings over time, they become popular or fade into obscurity. Because of its importance, language is studied by academics and public alike, as shown by the large number of publications and websites devoted to language evolution, etymology and semantic changes (Cresswell, 2010; Ayto, 2011; Lewis, 2013). Most of these focus on individual words only or are done on a small scale, mainly because the analysis requires manual work to locate occurrences of features in old texts, and then to compare manually their contexts or other characteristics.

In the recent years, large amounts of digitized old books and texts were made available, such as Google's Books initiative (Michel et al., 2010) with 5% of books ever published. Computational approaches have also been conducted to analyze them (Gulordava and Baroni, 2011), proposing novel approaches for understanding lexical semantic change – for an overview, we refer to the survey by (Tahmasebi et al., 2018). However, to the best of our knowledge, no cross-language temporal analysis has been proposed in the literature using computational approaches and large data. In addition, most prior studies focused only on English, whereas comparing two or more languages can shed light on how they actually co-evolved over time.

To study multiple languages over time, we assume the most intuitive approach: we focus on their similar connecting aspects. We use in particular words in both languages that have the same origins and similar meaning, also known as cognate words. We propose to study the temporal characteristics of cognate words as an approach to cross-language diachronic analysis. These cognates, loanwords included (i.e., words that come directly from the other languages) are an important subset of the lexicon and have been frequently studied. Most prior works focused on synchronous analysis of cognates (see for example (Uban et al., 2019)), while we look at their temporal aspects and correlations.

We have used the largest multilingual corpora available on a relatively long time, allowing thanks to its size to set a yearly granularity of analysis. In particular, we used Google Books Ngrams[1] in English and French to conduct the analysis. Despite its inherent problems (Pechenick et al., 2015), it is one of the few corpora of this size available in both French and English. We also prepared a list of English-French cognates based on existing lists and few selection criteria described below.

Cognates are, in linguistics, words that share a common etymological origin (Crystal, 2011), of which loanwords (words borrowed from other languages, e.g. English `communiqué` is borrowed from the French) are particular cases. Both are of great interest in multi-language analysis thanks to the ease of understanding and the identification of links between languages.

Numerous works have focused on either cognates or loanwords. On the one hand there are works for cognate detection harnessing computational methods that propose the first step in a (semi-) automatic analysis of cognates using the vast amount of digitally available data, when manual annotation requires a lot of man-hours (Jäger et al., 2017; List et al., 2018). On the other hand there are semantic analyses of cognates, that manually investigate cognates to look for links between two different languages (List et al., 2018; Aske, 2015). Some recent works cope with the limitations of these two categories by mixing the use of automatic detection of cognates with the semantic analysis (List et al., 2018; Rabinovich et al., 2018).

Nevertheless, to the best of our knowledge, there has been no automatic study of the frequency correlations and patterns of cognates over time across different languages, especially one that uses large datasets. In this paper, we propose a statistical change-oriented analysis of cognates, and focus on English and French.

[1] `https://books.google.com/ngrams`, accessed on November 15, 2019

## 2. Datasets

We started the study of English-French cognate by constructing a large cognate dataset that fits our criteria (see Section 2.1.). First, we created a list of cognates applicable for our study, basing our selection on available English and French lists of cognates (Bergsma and Kondrak, 2007), removing those that did not fit our criteria and adding some other. Each word's "cognateness" was confirmed by investigating its etymology with the Oxford English Dictionary, the on-line etymology dictionary[2] and the French National Center for Textual and Lexical Resources (FR: *Centre National de Ressources Textuelles et Lexicales*).

We used the 1-gram from the Google Books n-grams, for English and French (Michel et al., 2010) as an underlying dataset. It contains around half a trillion English words and one hundred billion French ones coming from books of varying literature genres. We note that although the dataset is not balanced in terms of document types its strong advantage lies in the very large size in comparison to other similar datasets, both in number of words and periods covered (from the 1500s to the late 2000s).

Finally, we would like to mention that we first focus on the differences in use frequency of words over time, hence we chose Google Books 1-grams. However, the underlying dataset can be easily extended by using larger n-grams such as 5-grams.

### 2.1. Criteria for Selecting Words

We chose English-French word pairs for constructing the cognates dataset and we based the selection on four criteria as follow. (1) We restricted the time scope to the years from 1800 to 2008, where most of the data is. (2) We chose words that were cognate pairs based on their etymology to make sure they were actual cognates. (3) We discarded verbs as their many inflections in French introduce noise, mostly as shared surface forms with other lexical items. (4) Finally, we chose words that appeared above a minimal frequency threshold (one in two million, or from 35 to 10,000 appearances in a single year, depending on the number of words available for that year) in both English and French to allow a proper analysis and to minimize the chance of an erroneous detection.

Once all words were selected, every inflection of each word was found using dedicated dictionaries. The frequency of all forms of a word were summed for each year to compute the total frequency of the word for that year. We then obtained for each word a time series from 1800 to 2008 representing its frequency. Finally, for each word, the time series, year of the first appearance, the maximum frequency and its year are all stored in a text file.

### 2.2. Cognates Dataset

Based on the data and the criteria presented above, we built, and release, a cognate dataset with 492 word pairs composed of nouns, adjectives and adverbs[3]. Each pair has between one and four forms in English, and up to ten in French. In English, most words have only one form for adjectives and adverbs, while most nouns have two forms (singular and plural). In French, with masculine and feminine, singular and plural forms, most nouns and adjectives can be found in four different surface forms.

The dataset includes 353 (71%) French loanwords (French words used in English) and 15 (3%) English loanwords[4]. These numbers include words taken from Old French and Old English. Note that the words are eclectic, both in meaning, as we aimed not to bias the dataset to any topic, and in frequency, as shown in Figure 1 where we plot median frequency as well as quartiles.

In the end, the dataset contains, for each cognate, both in English and in French, its frequency all inflexions combined in each year from 1800 to 2008 (0 in years before they appear or they are not part of the dataset).

## 3. Temporal Analysis of Cognates

We present below the preliminary results of the frequency analysis using the constructed cognate dataset.

### 3.1. Correlation of Cognates

First, we wanted to examine if the level of use of words in each of the languages changed in their own way or, rather, if the cognates shared similar patterns of changes in the intensity of their use over time. We then started by computing the frequency correlation for each pair of cognates. We used Pearson correlation coefficient(Pearson, 1895) on the time series representing cognate use in the concerned period. The frequency of a term in a given year is computed by dividing the number of occurrences of the term (the sum of the number of occurrences of each of its forms) by the total number of summed appearances of all words in this year.

As shown on Figure 2, there was a strong positive correlation for most pairs, with more than half (57%, 281) having a correlation value above 0.5, and over 13% (65) above 0.9. However, the high positive correlation is not true for every pair, as correlations go from -0.87 for the pair `employee - employé` to 0.99 for the pair `traditionally - traditionnellement`. Nevertheless, the number of pairs with a negative correlation, or close to zero, is rather small, as shown on Figure 2. This suggests that *cognates do not only share a past (etymological roots), but they also share similar usage patterns over time*.

Most of the cognate pairs had correlated changes of frequency over time. On the left of Figure 2, negatively correlated words are quite rare (6%, 31 words below -0.3). This suggests that cases when cognate words have tendencies to change the frequency of their use in an opposite way are quite rare.

If we restrict the analysis to the French loanwords (see the red plot in Figure 2), the positive correlation is similar, 201 loanwords (57%) having a correlation above 0.5 with their counterpart and 46 (13%) having the correlation value above 0.9.

[2] Available online at `https://www.etymonline.com/`

[3] Available online at `https://zenodo.org/record/3688087`.

[4] Due to the small number of English loanwords, we will focus only on French loanwords in our analysis.

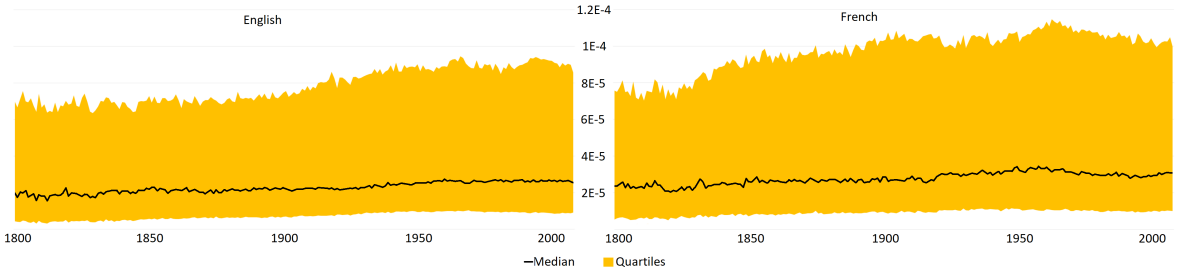



Figure 1: Distribution of the frequencies of cognates pairs, expressed through the quartiles and median.

### 3.2. Level of Word Use

The correlation of fluctuations in word frequencies over time as studied above still does not tell us whether words were actually used at the similar intensity levels in the same years. One word in a cognate pair could be used very frequently, while its counterpart could be barely used even though their relative frequency changes over time may be correlated.

To compare whether the frequency of a word is similar to its cognate counterpart, we first looked at the ratio between their maximal and mean frequencies. Then, for a cognate pair $(w_E, w_F)$, with $f_E(w, y)$ and $f_F(w, y)$ denoting the frequency (respectively, in English and French) of the word $w$ in year $y$, we computed the following formula:

$$\frac{max(max_{y\in[1800;2008]} f_E(w_E, y), max_{y\in[1800;2008]} f_F(w_F, y))}{min(max_{y\in[1800;2008]} f_E(w_E, y), max_{y\in[1800;2008]} f_F(w_F, y))}$$

This equation gives a real number of one or greater and is based on the comparison of the maximum frequencies of cognates. The closer to one, the greater the similarity between the maximum frequencies of the two cognates, with the limit at one where both the values (maximum frequency in English and maximum frequency in French between 1800 and 2008) being equal. When the resulting value is higher, the two words in a given cognate pair have a less similar use.

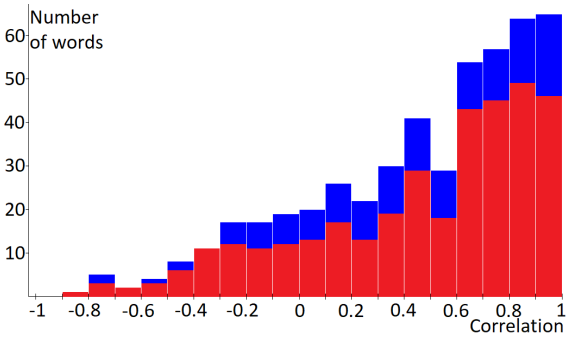



Figure 2: Correlation of English-French cognate pairs (blue) and French loanwords (red), from the first appearance of a word (English or French, depending on the earliest one) to 2008, as including earlier years would artificially increase correlation.

The cognate words not only tend to be correlated in terms of their changes over time, but they also have (for most of them) a similar level of use in their languages. The maximum usage of the most used word in each cognate pair is, for more than half of the words, at most 1.63 times more than its counterpart in the other language.

Moreover, the more we focus on the correlated words, the smaller this median line is (1.53 for correlation above 0.5; 1.49 for correlation above 0.7; 1.48 for correlation above 0.9). If we analyze only the loanwords, the results are similar.

To see if this ratio changes according to the frequency in one or both languages, and if one language has the cognates consistently more used (especially interesting are outliers), their respective mean frequencies seem to follow a linear distribution (see Figure 3). However, there are also cases of high frequency of use of a cognate in one language with low frequency in the other language (even several thousand times more in one language).

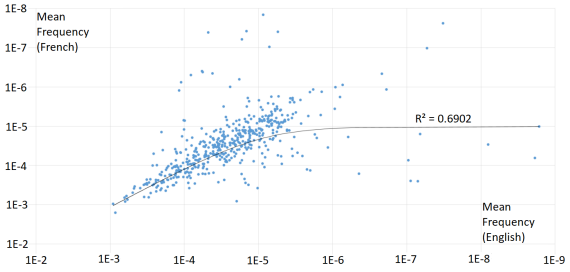



Figure 3: Distribution of the mean frequency in French according to the mean frequency in English (log-log plot). The linear regression $y = 1.1457x + 10^{-5}$ (black) shows the global relation between mean frequencies.

These extremes tend to be as likely to result from higher use in English as in French. As the correlation analysis indicated that the level of use of cognates evolved according to the same pattern across time, the frequency ratio indicates the *cognates have a similar level of use in both languages across time*.

### 3.3. Language Specificities

As the results show that cognate words are often used similarly at the same time in both the languages, one could be

tempted to say that a cognate, independently of language, performs in general a similar role in both languages and is used in very similar ways over time.
There are several potential reasons that could be proposed behind the differences in use frequencies and their temporal variations over time in both languages. To a certain degree, these could be explained by the subtle differences in the meaning of the cognates in both the languages, which would be used for slightly different purposes or in differing situations. Another driving force behind the observed differences in cognate use could be the existence of a synonym or multiple synonyms in only one of the two languages, which could "drain" the usage of one of the two words of the cognate pair: as per (Saussure, 1916), there is no bijective relationship between words in different languages.
Another explanation could be the occurrence of an additional acquired sense behind a cognate in one language increasing the use of this word with relation to its use in the other language. For example `azote` is barely used in English, in favor of `nitrogen`, while it is the opposite in French (`nitrogène` exists, yet `azote` is more commonly used).

### 3.4. Impact of External Factors

French and English are not only affected by each other, but by a multitude of external factors which can explain at least some of the correlations between cognates pairs, like the common history of corresponding countries. Analyzing history – i.e., the context around language use – can lead to an understanding of the impact of important events on some words, the most explicit example in our dataset being `bombardment - bombardement`, shown in Figure 4, a word which was obviously used more frequently in times of war, or, rather in the case of our corpus, when war-related books were popular. However, such effects are often difficult to determine, especially when the causes are less known.

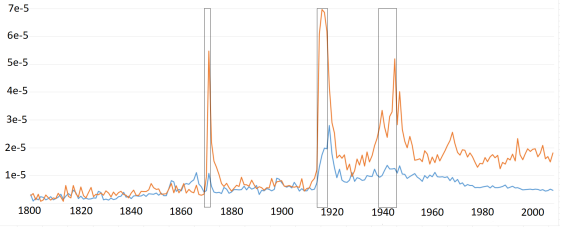



Figure 4: Frequency of `Bombardment` (English, in blue) and `Bombardement` (French, in orange) from 1800 to 2008. Three spikes can be observed (denoted by black rectangles), which correspond to the Franco-Prussian war (1870-1871), World War I (1914-1918) and World War II (1939-1945), showing the effect of the events on the languages.

## 4. Limitations

The dataset is not exempt from limitations, from its rather small size, as we focused on most-known cognates for the first analysis, to potential bias coming from the choice of words, even if we did our best to limit it, or from the corpus choice. We also provide the results of preliminary frequency-focused analysis of the cognates based on the created dataset. The analysis itself has some limitations: as it only covers two well-known languages, English and French, and only by not taking into accounts synonyms that made some cognates out of use in one of the two languages.

## 5. Conclusions & Future Work

In this paper, we describe a dataset of English and French cognates constructed to study their evolution from 1800 to 2008.
Diachronic language analysis and in particular studies of word origins have recently attracted considerable attention. In this paper we also emphasized the idea of studying temporal variability of a language by its synchronized comparison with another language where the synchronization is based on using cognates (serving as a comparative "bridges") aligned over time. By this, we add a second dimension or an additional investigation axis to the usual diachronic analysis approaches.
In the future, we plan to extend the current study to embrace larger number of cognates and to conduct a semantic analysis of the cognate variation across time and languages. We will also study other language pairs including ones that had less interaction and exchange in the past.

## 6. Acknowledgments

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 825153 (Embeddia) and 770299 (NewsEye).

## 7. Bibliographical References

Aske, J. (2015). Spanish-English cognates: An introduction to Spanish linguistics. Open Access eBook (Open Textbook). CC BY-NC-ND 3.0 US. (version: 29 June 2018).

Ayto, J. (2011). *Dictionary of Word Origins: The Histories of More Than 8,000 English-Language Words*. Arcade Publishing.

Bergsma, S. and Kondrak, G. (2007). Alignment-based discriminative string similarity. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 656–663.

Cresswell, J. (2010). *Oxford Dictionary of Word Origins*. Oxford University Press.

David Crystal, editor. (2011). *A Dictionary of Linguistics and Phonetics (6th ed.)*. David Blackwell Publishing. p. 104, ISBN 978-1-4443-5675-5. OCLC 899159900.

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.

Jäger, G., List, J.-M., and Sofroniev, P. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the*

*European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain, April. Association for Computational Linguistics.

Lewis, D. (2013). *Now I Know: The Revealing Stories Behind the World's Most Interesting Facts*. Adams Media.

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., and Forkel, R. (2018). Sequence Comparison in Computational Historical Linguistics Phonetic Alignments and Cognate Detection with LingPy 2.6. *Journal of Language Evolution*.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2010). Quantitative analysis of culture using millions of digitized books. *Science*.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*, pages 240–242.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.

Rabinovich, E., Tsvetkov, Y., and Wintner, S. (2018). Native language cognate effects on second language lexical choice. *CoRR*, abs/1805.09590.

Saussure, F. d. (1916). Cours de linguistique générale, ed. *C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot*.

Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.

Uban, A., Ciobanu, A. M., and Dinu, L. P. (2019). Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166.