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/

gs of the Workshop on Language technology for digital historical archives with a special focus on Central-, (South-)Eastern Europe, Middle East and North Africa, pa
Varna, Bulagria,September 5, 2019.
https://doi.org/10.18653/v1/P17

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.

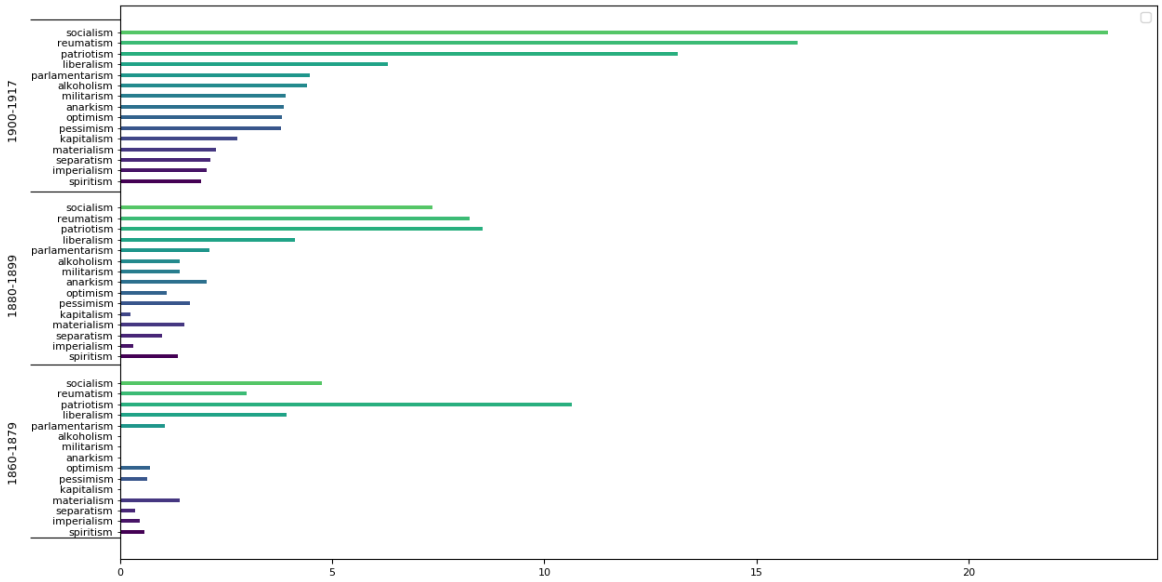
Both corpora are lowercased and lemmatized using LAS, an open-source language-analysis tool (Mäkelä, 2016).[2] LAS is a meta-analysis tool that provides a wrapper for many existing tools developed for specific tasks and languages. Though LAS supports multiple languages, most efforts were done to process 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

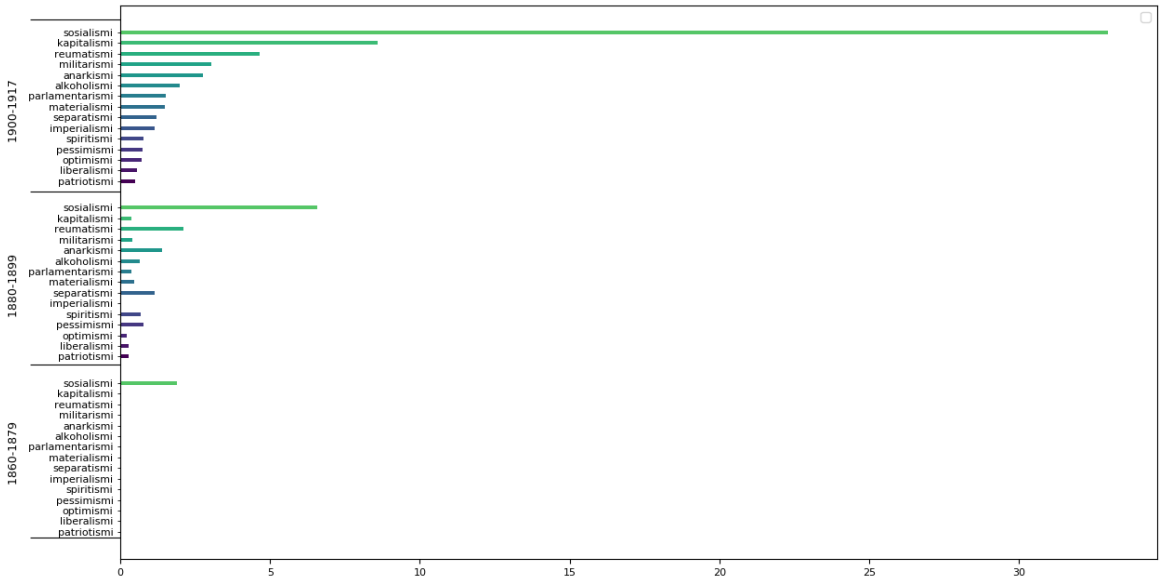
3.1 Diachronic embeddings

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

reumatismi ‘rheumatism’ *hiustauti* ‘hair disease’ *jäsensärky* ‘limb ache’ *hermo* ‘nerve’ *oxygeno* ‘?’ *vatsakatar* ‘gastritis’ *umpitauti* ‘constipation’ *nuha* ‘rhinitis’ *hermotautinen* ‘neurotic’ *topioli* ‘?’ *kurkkukatarri* ‘pharyngitis’ *parannuskeino* ‘remedy’ *hoitokeino* ‘cure’ *spirosiini* ‘spirosin’ *lazarol* ‘lazarol’ *lääkitä* ‘to medicate’ *kotilääke* ‘home medicine’ *reumaattinen* ‘rheumatic’ *hammastauti* ‘tooth disease’ *rautaliuos* ‘iron care’ *jäsenkolotus* ‘limb ache’ *leini* ‘rheumatism’ *linjamentti* ‘ointment’ *parannusaine* ‘betterment’ *vilustuminen* ‘cold’ *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'	*spiritism* 'spiritism' *teosofi* 'theosophy'	*spiritism* 'spiritism' *hypnotism* 'hypnotism'
pressensé 'presence' (Fr)	*frimureri* 'freemasonry' *feder* '?'	*andevärld* 'spirit world' *teosofisk* 'theosophic'
pater 'pater' *voltaire* 'Voltaire'	*mysterium* 'mystery' *spiritualism* 'spiritualism'	*spiritistisk* 'spiritualistic' *telepati* 'telepathy'
darwin 'Darwin' *renan* 'Renan'	*darwinism* 'darwinism' *positivism* 'positivism'	*själavandring* 'transmigration'
zola 'Zola' *newton* 'Newton'	*buddism* 'Buddhism' *darvinism* 'darvinism'	*trolleri* 'magic' *journalism* 'journalism'
balzac 'Balzac' *michelet* 'Michelet'	*vegetarianism* 'vegetarianism' *astrologi* 'astrology'	*ockult* 'occult' *astrologisk* 'astrological'
galilei 'Galileo' *corneille* 'Corneille'	*teosofisk* 'theosophic' *bibelkritik* 'Bible criticism'	*astrologi* 'astrology' *frimureri* 'freemasonry'
aristoteles 'Aristotle' *kabal* 'cabal'	*metafysik* 'metaphysics' *teosofien* 'theosophy'	*gondiagnos* 'eye diagnosis' *alkemi* 'alchemy'
oppert 'Oppert' *rousseau* 'Rousseau'	*darvin* 'Darvin' *darvins* 'Darvin'	*clairvoyance* 'clairvoyance'(Fr)
proudhon 'Proudhon' *zolas* 'Zola'	*utvecklingslära* 'evolution' *malthus* 'Malthus'	*tankeläsning* 'mind reading'
quand 'when' (Fr) *loyson* 'Loyson'	*själavandring* 'transmigration'	*tungomlstalande* 'tongues'

Table 4: Clusters containing Swedish word *spiritism*.

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

Table 5: Swedish clusters containing word *separatism*

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

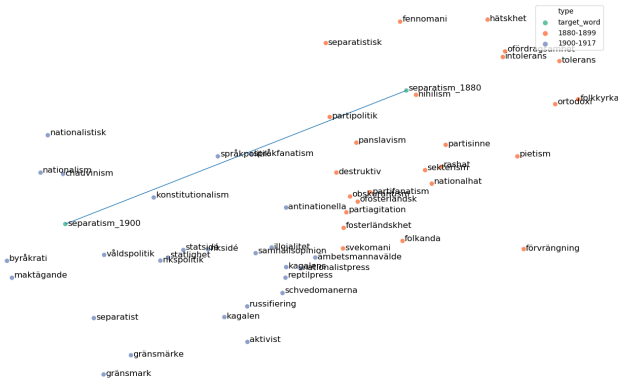
Table 6: Finnish clusters containing word *separatismi*

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.

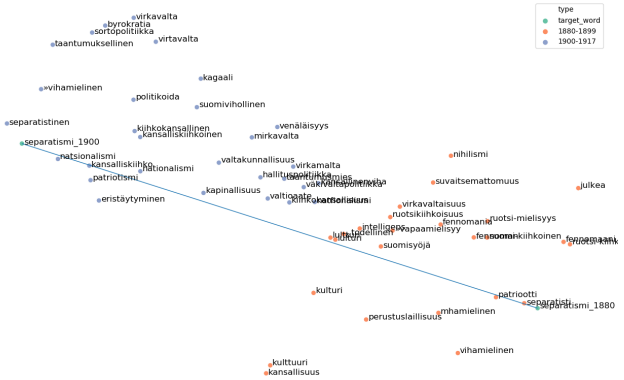
The Finnish clusters for *separatismi*, presented in Table 6, are quite similar to Swedish. The main difference is that in the 1860-1879 the word is mentioned less than 100 times and as a consequence excluded from our models. But the 1880-1899 and 1900-1917 Finnish clusters follow the same pattern: the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