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.

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 [**?**]. *Isms* are commonly associated with ideologies in modern political language, but not all *isms* are ideologies and vice versa [**?**]. In fact the relationship between *isms* and notions of ideology have changed historically and has varied depending on cultural context [**?**].

While there are works that study *isms* from a long-term perspective [**?**], their applicability as analytical tools for historiography [**?**], their rhetorical appeal [**?**], or their cultural transferability and special place in Chinese political language [**?**], 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 [**?**, **?**]. 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 [**?**]. While this paper does not assume one to one correlation between distribution and semantics, it uses word embeddings 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 [**?**, **?**]. 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 [**?**]. Though the archive contains newspapers starting from 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

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

languages. In our analysis, these dominant languages are treated as two separate corpora even though contemporaries often relied on newspapers in both languages [**?**]. 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 [**?**].[1] 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 [**?**] on each time slice. We use the Gensim Word2Vec implementation [**?**] 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 [**?**]: 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

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

from diverging rapidly. This approach has been previously used in [**?**] 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 [**?**]. 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 [**?**], with default parameters.

Affinity Propagation has been previously used for various language analysis tasks, including collocation clustering into semantically related classes [**?**] and unsupervised word sense induction [**?**]. 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[2]. 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.

[2]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 [**?**]. 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

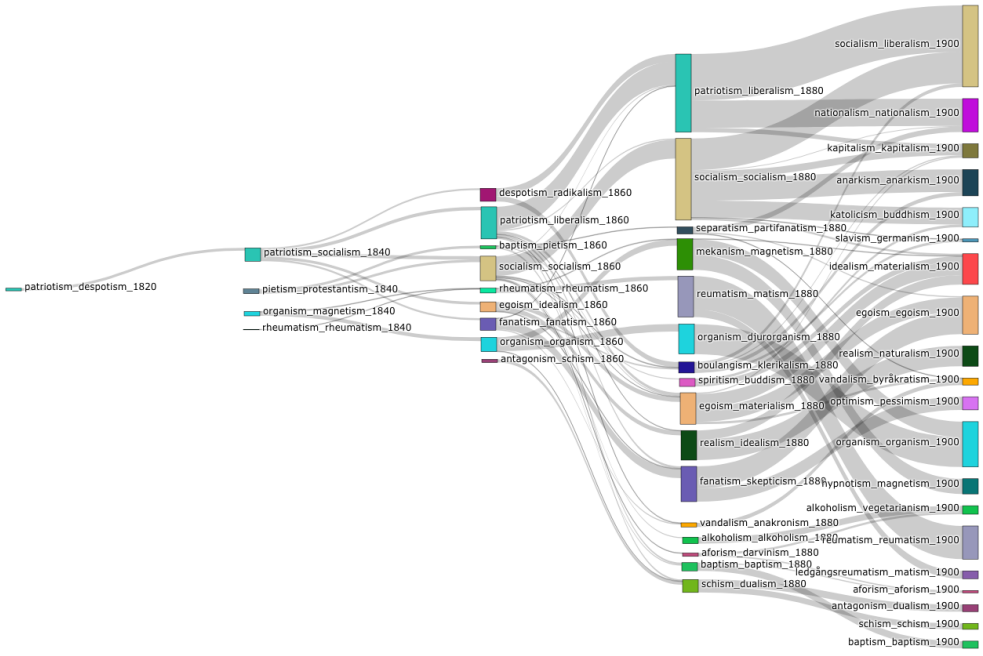
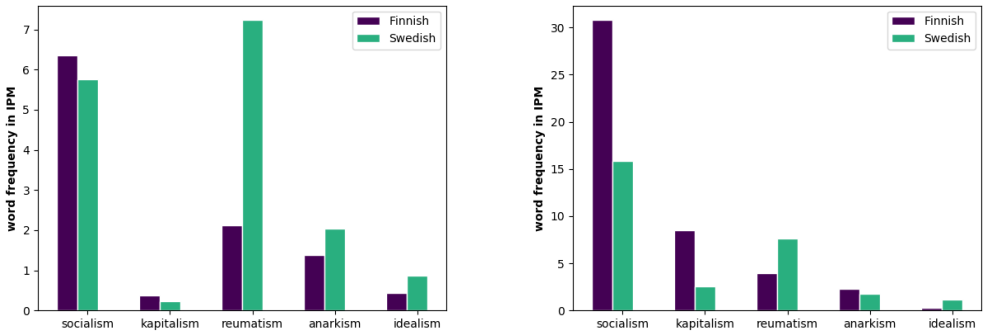


Figure 1: Sankey diagram of *ism* clusters from the Swedish dataset covering five double decades from 1820 to 1900. The cluster name is the most frequent ism word for that cluster followed by the cluster representative and the double decade.

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 [**?**]. 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.[3]

4 Discussion and Future Work

There are alternative ways to build diachronic embeddings. The recent line of research is aimed at smooth time representation [**?**, **?**, **?**, **?**]. 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.

[3] 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).