Studying the Historical Semantics of Finnishness with a Bigram Approach

Marjanen, Jani

CEUR-WS.org
2022

Marjanen , J , Kanner , A & Mäkelä , E 2022 , Studying the Historical Semantics of Finnishness with a Bigram Approach . in K Berglund , M La Mela & I Zwart (eds) , Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022) . CEUR Workshop Proceedings , vol. 3232 , CEUR-WS.org , Aachen , pp. 109-119 , Digital Humanities in the Nordic and Baltic Countries Conference , Uppsala , Sweden , 15/03/2022 .

http://hdl.handle.net/10138/350324

Studying the Historical Semantics of Finnishness with a Bigram Approach

Jani Marjanen[1], Antti Kanner[1] and Eetu Mäkelä[1]

[1] *University of Helsinki, Finland*

Abstract

Our paper analyzes the historical understanding of Finland and Finnishness as it was expressed in newspapers published in the late eighteenth century and the early nineteenth century. As the period saw the decimation of the Swedish Kingdom and establishment of the Grand Duchy of Finland within the Russian Empire, a change in language use can be expected, but the changes occurring are rather fine-grained and difficult to detect without a systematic and transparent charting of the data. This paper suggests a method based on the analysis of bigrams to study this type of semantic change. Many existing methods are designed to navigate massive amounts of linguistic data and do well in solving computational tasks, but are not always a good match for the kind of historiographical questions such as ours. More broadly, we establish that the application of Finnish can in principle refer to the categories of language, geography, nationhood, statehood, and territoriality. Our analysis shows that especially the categories of language and state underwent a gradual shift in the period from the eighteenth to the nineteenth century.

Keywords

Ethnonyms, conceptual change, historical newspapers, bigrams

1. Introduction

The year 1809 is sometimes referred to as an “annus mirabilis” in Finnish history [1], as it effectively meant the establishment of a Grand Duchy of Finland within the Russian Empire and paved the way for state and nation building. How Finland and Finnishness was understood, clearly changed in the process, but how to describe this transformation has been heavily debated – at least since Zacharias Topelius’s famous essay “Do the Finnish people have a history?” from 1843 [2]. Previous scholarship has either emphasized that there are clear continuities from how Finnishness was understood in the eighteenth century to the nineteenth century or stressed the rupture that came about in 1809 [3]. The former line of interpretation has in general continued a form of national history writing that was born in the late nineteenth century and should in general be seen as a form of nation building in the Grand Duchy. The latter tradition is dominated by scholarship in the 1960s onwards that criticized the national perspective and sought to emphasize Finland’s imperial connections as well as highlight the constructed character of the nation. By concentrating on the notion of Finnishness as it was invoked through language use, this paper contributes to that debate.

By asking what kinds of things could be attributed as Finnish and what was meant with the Finnishness of those things, the study seeks to trace the broad patterns of language use regarding Finnishness before and after 1809. In doing so, the paper is couched in the latter historiographical tradition, but it also suggests that one of the problems in the debate has had to do with the fact that neither tradition has been very good at discerning how language use regarding Finnishness actually changed. Operationalizing this question as a corpus-driven study, allows for concentrating on the distribution and polysemy of “Finnish” and hence contributes to the historiographical debate.

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.
EMAIL: jani.marjanen@helsinki.fi (A. 1); antti.kanner@helsinki.fi (A. 2); eetu.makela@helsinki.fi (A. 3)
ORCID: 0000-0002-3085-4862 (A. 1); 0000-0002-0782-1923 (A. 2); 0000-0002-8366-8414 (A. 3)

Apart from participating in the debate in historical scholarship, the paper also contributes to a discussion about the aptness of various NLP methods in the digital humanities. With the rise of digitized historical corpora, new methods in NLP have increasingly been harnessed to study the semantic and lexical changes in long-term diachronic corpora [4]. Many of these methods are designed to navigate massive amounts of linguistic data and do well in solving computational tasks, but are not always a good match for the kind of historiographical questions such as ours. In the period we are interested in, the publication of newspapers and books in Finland remained at a modest level. It is in this regime of limited data that we make our methodological contribution, which is to apply a simple bigram approach to the analysis of a specific case of historical semantic change. By its general outline, our study falls under a wide definition of corpus linguistic studies in lexical semantics. Similar, corpus-driven studies on adjective groups with synonymous or otherwise similar meanings include a study on *powerful* vs *strong* by Church et al. [5], and a study on *chief*, *principal*, *major*, *main* and *primary* by Lin [6], among others. From the mainstream of corpus linguistics, our study differs not so much in methodology, but in that it seeks to target a very specific, historiographically informed interest of knowledge, whereas in corpus linguistics it is more commonplace to apply a data-driven perspective and base the distributional characterizations to the most statistically robust distinctions, regardless of what they are.

The major benefit of using such simple models in this regime is transparency. To answer our focused interest, it is imperative that the method used make possible transparently following and qualitatively understanding the process from the aggregates to the source material, so that 1) it is possible to delve into even individual examples in the source data for interpretation if needed and 2) the interpretation can combine information derived from various levels of aggregation within the method, as well as coming in from external scholarly historical understanding. Second, as the study is interested in semantic phenomena which can be adequately targeted by bigram-range environments, using richer distributional characterizations and more complicated algorithms to handle them, is unnecessary and from this perspective can only function as a source of noise. In the case of adjectives, bigram-range mostly captures cases when they are modifying head nouns. While there are other syntactic possibilities for adjectives to occur, it is fair to assume that in Swedish, the direct modifier of a nominal head is the prototypical occurrence type of any adjective. Positions, such as predicate adjectives or less typical modifying constructions such as coordinated modifiers are more marked options. As our study seeks to target what is most ordinary and commonplace in language use, trading away the less prototypical cases for more transparency is well founded.

Hence, our study concentrates on the meanings of the word *finsk* (“Finnish”) and other comparable adjectives, *svensk*, *rysk* and *skånsk* (“Swedish”, “Russian” and “Scanian”, respectively). It assumes that tracing the casual and established patterns of use are indicative of common, widely shared features of the concepts corresponding to these adjectives. Established lexicographical characterisations of these words show that they, like their English equivalents, are polysemous over ethnicity, geographic origin, language, citizenship and so on [7]. We maintain that even though the distinctions between these senses are relatively clear in out-of-context inspection, they are often un-analyzed in actual cases of use, especially in historical context. Many of our examples remain ambiguous over several of its senses even after close-reading. This is predictable from the perspective of semantic theories which project polysemous structures in terms of prototype categories, such as cognitive linguistics, where token-level meanings fall under radial categories formed around prototypical uses [8, 9]. Such theoretical underpinnings also predict a high degree of interference between different senses if their prototypes reside close by. Thus, inspecting the relative frequencies of typical use cases leads to observations about the prototypes of Finnishness, Swedishness and the like.

To aid our analysis, we establish that the application of *Finnish* can in principle refer to the categories of language, geography, nationhood, statehood, and territoriality. Comparing the categories before and after 1809 indicate that the role of language diminished, whereas the role of the state grew over time. There are also clear continuities that are discernible in the data, but the shifts we see in the bigrams clearly indicates how the language of Finnishness expanded to new domains of life after 1809.

Although powerful in many applications, the weakness in machine-learning methods often lies in the difficulty of connecting analysis results with concrete examples in historical sources. They are often based on convoluted abstractions of distributional tendencies of studied linguistic structures. While the result of our study is in itself rather intuitive – as it is clear that notions of Finnishness changed due to the establishment of the Grand Duchy of Finland – we believe that this also makes the strength of the method visible. Analyzing bigrams enables quantifying results in an understandable, transparent way, but also readily caters for studying the occurrences qualitatively in their contexts of use, that is, combining quantitative and qualitative analysis in the context of one research case. This is an important asset in cases where datasets are of smaller scale and the studied phenomena shows ambiguities when explored on type level.

2. Data

Our study draws on digitized historical newspapers from Finland and Sweden. The newspapers have been consulted in three different versions: the graphic user interfaces provided by the National Library of Finland [10] and National Library of Sweden respectively, the morpho-syntactically parsed versions of the data provided by the Language banks in Finland and Sweden [11], as well as a version of the newspapers *Tidningar Utgifne af et Sällskap i Åbo* and *Åbo Underrättelser* that has undergone an improved optical character recognition within the NewsEye project [12]. Our analysis of these two newspapers are based on the last version.

As our initial interest lies in the Finnish conceptualization of Finnishness in the period before 1809, we focus on the only newspaper published in the Finnish part of Sweden, *Tidningar Utgifne af et Sällskap i Åbo.* It was published in the largest city, Turku (Åbo), from 1771 onwards and was in effect the only newspaper in that town in the whole period up to 1809. As the paper also went through two small hiatuses and two name changes in this period [13], we have mainly focussed on its first period of activity from 1771 to 1779. To contrast our findings, we have selected two similar newspapers for additional analysis, *Lunds Weckoblad* published as a similar academically oriented newspaper in the Southernmost part of Sweden, Scania (Skåne) and *Åbo Underrättelser* published in Turku in the 1830s, so in the period after 1809. In choosing contrasting newspapers, we have sought after regional papers with a similar profile as *Tidningar Utgifne af et Sällskap i Åbo*. As newspapers at this time were in general rather varying in scope and could easily be shut down because of lack of resources or material [14, 15], we believe it is important to choose comparable regional newspapers. When appropriate, we have also compared our findings in the individual newspapers to the overall data in the national newspaper collections in the same period.

3. Method

In summary, we study the adjective *finsk* ('Finnish') in comparison to other common, comparable adjective in our data and classify their head nouns. For example, entities closely related to statehood (such as *crown*, *realm* or *history*) often have *svensk* ('Swedish') or *rysk* ('Russian') as attributes but never *finsk* in the eighteenth-century. Contrasting head nouns this way is a viable method for teasing out conceptual distinctions between adjectives and/or their referents as the semantic compatibility is necessitated by the direct syntactic co-selection of head noun and its adjective modifier [16]. We then proceed to statistically establish the effect of a limited amount of data on the observations about excluded categories.

Most linguistic theories assume that the meaningfulness of utterances requires a certain level of conceptual compatibility between expressions bound together by syntactic relations. Katz & Fodor approached this compatibility with the help of the concept of selection restrictions [17]. In Cognitive Grammar, independent structures (such as nouns) must fit the role allocated to them in the image schema representing the semantic outline of the dependent structure (such as predicate structures and modifiers) [18]. Pustejovsky's Generative Lexicon proposes a mechanism of type coercion, by which semantically compatible affordances of meanings of linguistic expressions are co-selected to produce understandable utterances [19]. Kanner has identified syntax-driven restriction as one of the three major handles distributional operations have on meanings of linguistic expressions [16].

The method used in the present study is thus based on extracting all bigrams where the studied adjective is followed by a noun which it modifies, counting their frequencies and semantically classifying them on the type level. Because the OCR quality of the data is relatively poor, the lemmatization is not entirely reliable. To tackle this, each studied adjective was searched by using its around 20 most common OCR variants. These were identified by looking at strings that had a low

Levenshtein distance to the correct words and then evaluated manually for inclusion. This considerably increased the frequency for each studied lexeme. Likewise, many of the OCR variants of the head nouns were merged in the analysis stage.

The semantic analysis of the head nouns was based on four categories. While the adjectives studied here could be broadly called ethnonyms, closer examination reveals that they seldom designated an attribute related to ethnicities or peoples. Instead, in the politico-cultural space of the late eighteenth to early nineteenth-century Northern Europe, the same words were used to designate ethnicity, language, statehood, territory or geographical area in different proportions. Instead of projecting a structure of polysemy, where these different categories correspondent to type level senses of *finsk*, Generative Lexicon [19] assumes that on the type level different adjectives have differing conceptual affordances to make up meaningful compositions with their head nouns regarding the semantic structure of each. Thus, mapping the head-noun distributions gives a hint of these affordances of each adjective, and helps to characterize which affordances are more commonplace and typical and which are peripheral and uncommon. It is these distinctions in conceptual affordances that our study focused on, as, for example, the use of Finnish in adjectival modification related to statehood became possible only when *finsk* attained this semantic dimension. Thus, the most dominant categories by terms of frequency can be used as indicators of the most unmarked facets of Finnishness (or e.g. Swedishness). Thus, we do not propose that correspondence to below categories would be indicative of differing senses of *finsk*, rather they are our analytical heuristic in charting its conceptual affordances and their internal hierarchy.

- **State** – The modifying adjective designates the state to which the head noun belongs or is affiliated to. It distinguishes the head noun from similar entities belonging to other states. Examples most often include state institutions, such as armies or the crown.
- **Geographic** – The category of geography implies that the head noun is a geographical entity residing in the area specified by the modifier or that the head noun is something that simply comes from that geographic area. In the latter case, the category can be difficult to distinguish from the categories of state and nation. Examples include calling lakes Finnish, but sometimes also groups of people who are linked to the geographic area (e.g. Finnish commoners).
- **Nation** – The modifying adjective implies a national or ethnic dimension of the head noun. This is evident in items such as “Finnish nation”, but also items that can be seen as belonging to the nation, such as “Swedish history”.
- **Language** – The modifying adjective designates the language of the head noun and the distinction it makes distinguishes the head noun from similar entities which are in different languages. Examples include documents and texts, such as bibles or lexicographies, but also of events like masses.
- **Territory** – The modifying adjective indicates the territory of the state. Examples include “the Swedish border” or “Swedish Pomerania”.

The analysis was conducted manually on the type level. If type-level classifications proved difficult due to ambiguity, the analysis included token level inspection. In some cases, this ambiguity was also present on the token level. However, the general quantitative patterns in proportions between the categories were satisfactorily robust, even when taking the few ambiguous bigrams into account.

4. Results

4.1. The period before 1809

In order to say something about the meaning of the word Finnish, we have to compare it with other similar words. The closest candidate is the word Swedish, which also makes sense as a point of comparison as *Tidningar Utgifne af et Sällskap i Åbo* was published in the Finnish part of the Swedish kingdom. Both words are fairly frequent in the newspaper, and many studies have indicated that intellectuals in Turku represented a kind of supplementary patriotism, in which they directed devotion to Sweden as a realm, but also Finland as a region. Improving conditions in Finland was a way of helping the whole Swedish realm. This sentiment is also present in the Aurora Society, the organization that produced *Tidningar Utgifne af et Sällskap i Åbo* [20–22]. Some scholarship has been eager to emphasize the Finnishness of the Aurora Society and its newspaper, even by seeing them as vehicles for burgeoning national awareness [23]. The newspapers position as Finnish, on the one hand, and Swedish, on the other hand, is also visible in the history writing of the newspaper. The newspaper has since the latter half of the nineteenth century been celebrated as the first newspaper published in Finland, giving it a position as a national symbol [24]. Professional historiography on the development media has been downplaying this role, by making it clear that *Tidningar Utgifne af et Sällskap i Åbo* was published in Finland, but should be seen as a part of the developing regional press in the Swedish kingdom [13, 25]. Against the backdrop of these interpretations that give different emphasis to the newspapers Finnishness and Swedishness, it is interesting to study the meanings present in the newspaper itself.

In the 1770s, the term Swedish was used slightly more often than the word for Finnish in the newspaper, but the roles were reversed in the 1790s. As an exploratory step, pairwise cosine similarities based on collocations (extracted from +/–5 words range and using mutual information MI as collocation metric) were computed between all words belonging to the comparable frequency band. Words for Swedish and Finnish were found to be their closest neighbors in this analysis and a considerable part of the similarity stemmed from their similar head-noun selections. While there are good linguistic grounds for concentrating on the use of these adjectives as modifiers, this exploration showed how this particular distributional feature was the one that seemed to set them apart from other words, including other adjectives of comparable frequency.

Still, the modifiers "Swedish" and "Finnish" shared only a few head nouns. They are "book", "city", "congregation", "burghers", and "name". Of these five, three are related to language as indicated in Figure 1, with the color yellow. In the phrases "Finnish city" or "Swedish city" the ethnonymic words refer to geography rather than anything else. The "burghers" is vague and hence more difficult to categorize according to our scheme. It refers to the burghers as an estate in the representative organ in the Swedish realm, so it definitely belongs to the state apparatus, but in the case of the Finnish burghers, it indicates the Finnish nation within the Swedish kingdom. The uncertainty in categorizing head nouns is there throughout the data, and while it can be done in most of the cases, there still is a level of uncertainty to any quantitative assessment of the bigrams according to the categories we use.

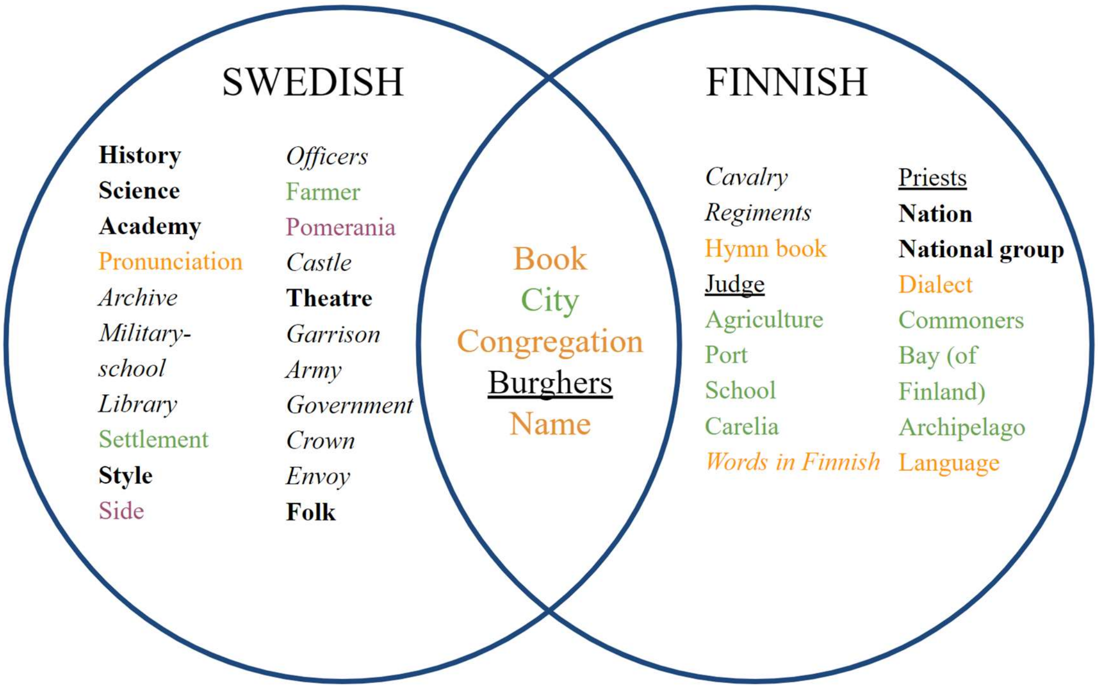

Figure 1: Diagram of the most prominent head nouns in the bigrams starting with either Swedish or Finnish in *Tidningar Utgifne af et Sällskap i Åbo*. Terms are categorized according to language (yellow), geography (green), state (italics), nation (bold), and territory (purple). The underlined words burghers and priests are ambiguous in their categorization, but refer to state and nation.

In general, the Swedish bigrams refer more often to statehood, nationhood and territory than the Finnish ones. The head nouns that are modified by "Finnish" overwhelmingly relate to geography or language. Of the classified nouns 44 percent are related to language and 33 percent to geography, whereas associations to the state and nation hold for 11 percent each (see Fig. 2). But, the quantification needs to be taken with a grain of salt, as a majority of nouns are impossible to classify due to their ambiguity. What is perhaps more revealing is that there are no nouns coupled with "Finnish" that can be reasonably classified as territorial. The nouns that relate to "state" are dominantly about the military and are included in historical accounts of Swedish war efforts in which particular regiments or cavalry troops are mentioned as "Finnish". In these cases the state oriented nouns are actually of the terms like the "Swedish army", so even in the cases in which Finnish is somehow related to statehood, it matches the earlier described supplementary character of the Finnish-Swedish configuration. In the eighteenth century, many of the war efforts specifically involved Finnish troops and some documents did mention the Finnish army [26].

Tidningar Ugifne af et Sällskap i Åbo writes both about Finnish and Swedish things, and the newspaper itself comes across as both Swedish and Finnish, but the words had slightly different emphasis. Finnishness was primarily geographical and linguistic with links to a Finnish nation in the sense that Finnish-speakers were clearly acknowledged as a separate group in the kingdom, people could be regarded as belonging to the Finnish nation, and the estates of priests and burghers could also represent Finland. Like Jonas Nordin has emphasized, the language question gave Finland a special role in the ethnic mix in early-modern Sweden. The fact that Sweden was constituted by several different parts that had their specific relationship to the crown, was visible in, for instance, the title of the monarch which in its abbreviated form explicitly mentioned Sveas, Götar and Wends, but in everyday language, as represented by newspapers, it was quite common to use the phrase "Sweden and Finland". This phrase did not mean that Finland was seen as less integrated into the realm, but that the linguistic and geographic distance provided a need for distinction [21]. That the term Swedish referred more often to statehood and territory (although the latter was fairly seldom) is in itself not surprising, but the difference is important to note. A comparison to the bigrams containing the word "Russian" highlights this, as those frequently contain nouns relating to state institutions and even territory.

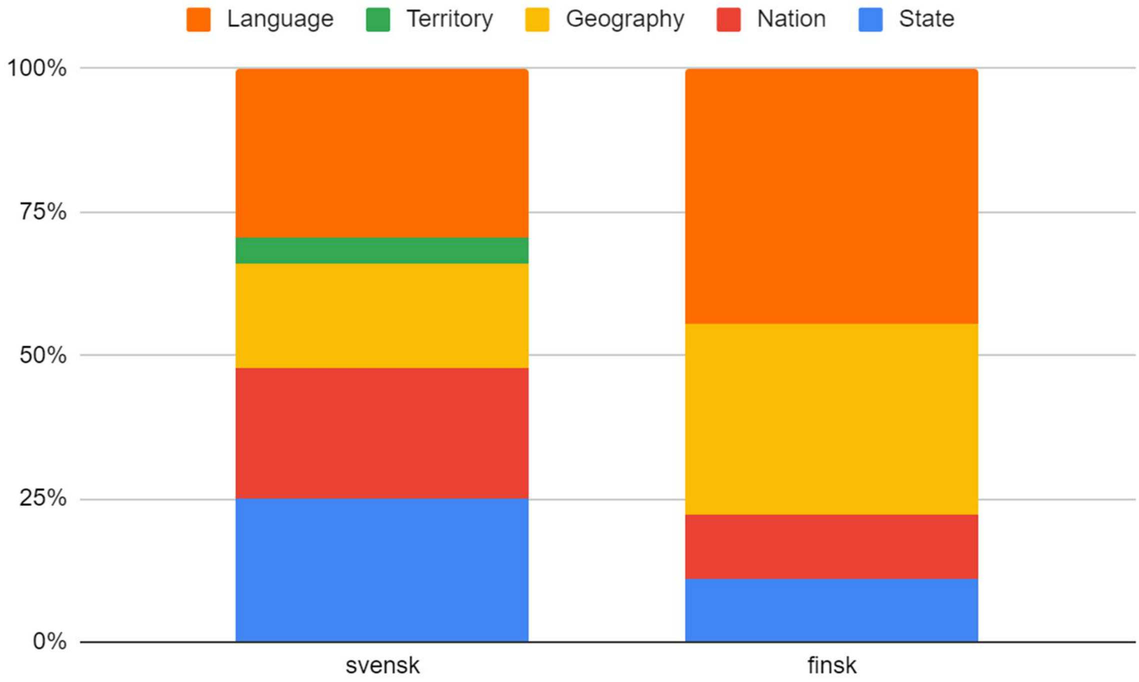

Figure 2: Domains of "svensk" and "finsk" in *Tidningar Utgifne af et Sällskap i Åbo* for 1771–1778. The language and geography domains dominate the uses of "finsk", whereas the uses of "svensk" are more evenly distributed.

To better understand the discourse of Finnishness in *Tidningar Utgifne af et Sällskap i Åbo*, further comparisons are needed. To do this we compared our findings with ethnonymic expressions from *Lunds Weckoblad*, a regional weekly published in Scania at roughly the same time. The newspaper from Southwestern Sweden did not write much about Finland. The term "Finnish" appears only four times, but quite interestingly the focus is very different. The mentions all relate to military campaigns with nouns such as "war", "army", "troops", and "border". The last one is difficult to classify, as we see it as a marker of territory, but read in context it can also be seen as a geographic marker indicating the part of the Swedish border that lies in Finland.

The more interesting comparison has to do with how *Lunds Weckoblad* used the terms "Swedish" and "Scanian" (*skånsk*) as compared to how *Tidningar Utgifne af et Sällskap i Åbo* wrote about "Swedish" and "Finnish". As another regional newspaper in the Swedish kingdom, it could also refer to Sweden as a whole, but also cover regional matters. The profile of the modifiers of "Swedish" is very similar. In *Lunds Weckoblad* the distribution in our classification is 23 % for state, 18 % for nation 25 % for geography, 4 % for territory, and 32 % for language, whereas the shares in *Tidningar Utgifne af et Sällskap i Åbo* are 25 %, 23 %, 18 %, 5 %, and 30 %. Bearing in mind some of the uncertainties in the classification, we refrain from drawing strong conclusions other than that the way of using the ethnonym Swedish in Scania and Finland was rather similar. Comparing the use of "Scanian" in *Lunds Weckoblad* and "Finnish" in *Tidningar Utgifne af et Sällskap i Åbo* does show some differences. "Scanian" could not be used to denote linguistic features, but its uses were also otherwise limited as more than half of the uses relate to geography and 40 % refer to different military organizations. For Finnish, the category of language is the largest one with 44 %, but also here geographical markers are among the most important ones with 33 %. In the region to region comparison, it is particularly the language aspect that sets Finland apart. It also seems that Finnish as a language was not written about in *Lunds Weckoblad*.

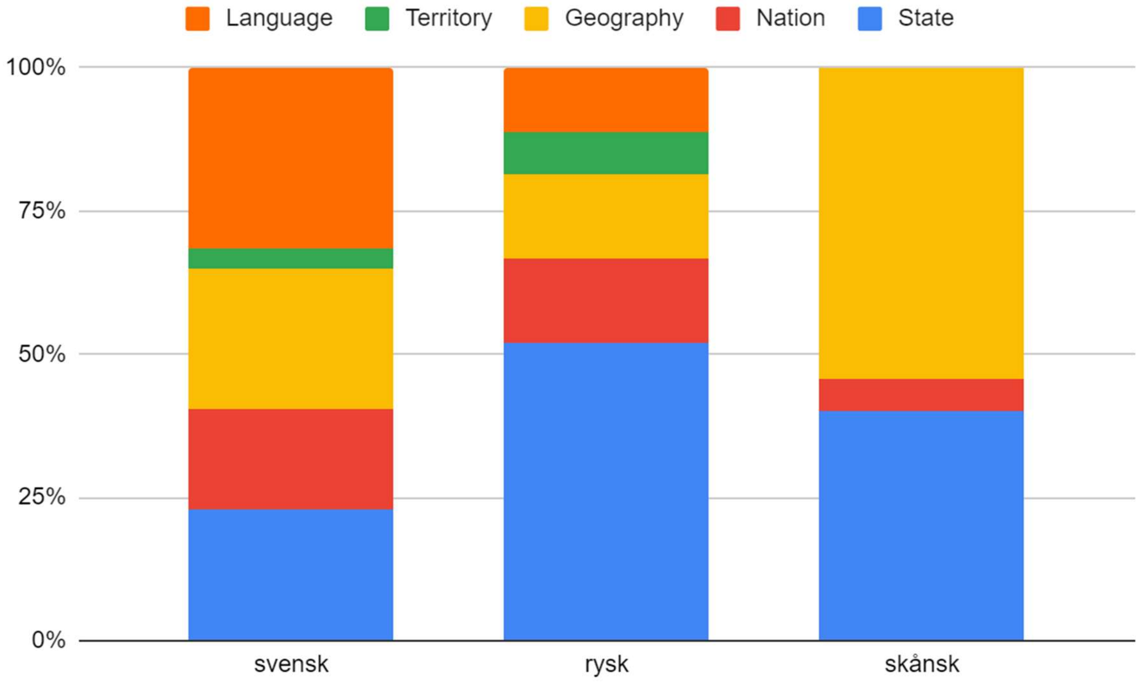

Figure 3: Domains of "svensk", "rysk" and "finsk" in *Lunds Weckoblad*. The profile of "svensk" resembles that of *Tidningar Utgifne af et Sällskap i Åbo*. The profile of "skånsk" in the Scanian newspaper resembles that of "finsk" in the Finnish newspapers, but does not include the emphasis on language.

Compared to the big papers in Stockholm, this is a big difference. By looking at all the newspapers published in Sweden at this time, the word "Finnish" occurs most often in the major Stockholm newspapers *Dagligt Allehanda* and *Inrikes Tidningar*. The four most frequent bigrams in them relate to language, indicating that the conceptualization of Finnishness in the political and cultural epicenter of the kingdom had predominantly to do with language.

However, the overwhelmingly most common bigrams relate to regular announcements of the Finnish-language masses in Stockholm, which means that the nature of the data also underlines this aspect. Striking a perfect balance is difficult. On the one hand, if our data would cover only uses of the word "Finnish" in prose text, and exclude advertisements or announcements, we would see fewer bigrams relating to language, but also to Finnish produce (such as butter or meat), which we have classified under geography. On the other hand, advertisements and announcements were an important part of what the newspapers communicated and hence also part of the historical record that is included in the non-standard data we use [27].

4.2. The period after 1809

To get a temporally usable comparison from the period Sweden had ceded Finland to Russia and the Grand Duchy of Finland was formed, we chose not to analyze papers directly after 1809, but instead chose to analyze *Åbo Underrättelser* from the 1830s. The reasons for this are twofold. First, we wanted to use a regional newspaper, and, second, we wanted to look at a period when the so-called divorce between Sweden and Finland was settled. By the 1830s, the Russo-Swedish alliance of 1812 had ended, and Finnish political activists had started talking about a Finnish state [28, 29]. Although there were some revanchist sentiments in Sweden (including some Scandinavianist ideas) [30, 31], the question of Finland becoming Swedish again had largely become a non-issue. It can also be assumed that contributors to local newspapers had by the 1830s a distance to the Swedish era.

In *Åbo Underrättelser* the terms Finnish, Swedish, and Russian appear all quite frequently. Including the erroneous extra bigrams produced by poor OCR, all three include more than 200 types of bigrams. By classifying them, we can however see that the ethnonyms have slightly different profiles. For the classification, we omitted unclassifiable words and obvious errors produced by insufficient OCR.

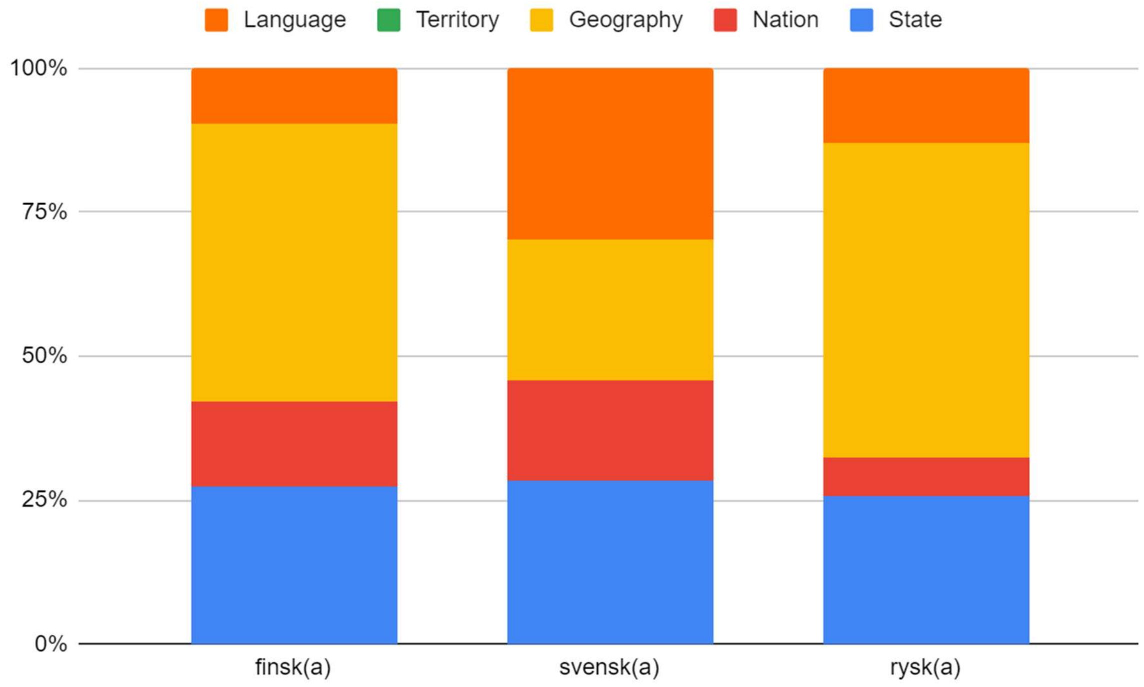

Figure 4: Domains "finsk", "svensk", and "ryska" in *Åbo Underrättelser* from the 1830s.

Again, there are some uncertainties with regard to the classification and the data quality, but some changes are obvious when compared to the eighteenth-century newspapers (Fig. 2 and 3). For a Swedish-language paper published in the Grand Duchy of Finland, matters that related to the Swedish and Russian state were not as prominently present as before. Proportionately, the themes of state and nation became more prominent in the data from the 1830s, but perhaps more importantly, the bigrams classified as relating to the state, became more diverse. Military terms dominated in *Tidningar Ugifne af et Sällskap i Åbo* and *Lunds Weckoblad*, but in *Åbo Underrättelser* we also encounter "bank" and "citizens" as terms in that category.

Another interesting theme is that, bigrams starting with Finnish are no longer as focussed on language as before. Instead, a slim majority of bigrams starting with Swedish had to do with language. This development points at a near reversal, in which Finnish language dominated the picture in eighteenth-century Swedish newspapers, but in *Åbo Underrättelser* from the nineteenth century Swedishness is most often connected to language. The reasons for this are not straightforward, but reading text samples suggests that it relates to the fact that Finnish-language newspapers had started appearing in Finland, making the issue of the Swedish language more pertinent in the Finnish context. In other words, when referring to the language dimension of Swedishness *Åbo Underrättelser* often referred to the Swedish language in Finland. In Stockholm in the eighteenth century, what set Finland apart the most was its majority language, Finnish, whereas this did not seem to be that important on the Eastern side of the Bay of Bothnia. After 1809, Finland's speciality vis-à-vis Stockholm was no longer about language, but about not being part of the kingdom, whereas in the newly established Grand Duchy, the language issue gradually became an internal matter for the Finnish nation.

A further explanation to the differences between *Tidningar Utgifne af et Sällskap i Åbo* and *Åbo Underrättelser* has to do with the changing position of Turku. In the eighteenth century, Turku was the second most important city in the kingdom and the intellectual center of Finland. In this position, it in a sense represented Finland toward Stockholm. The newspaper published by the Aurora Society was a crucial channel for this [3]. After 1809, this changed as Helsinki (Helsingfors) had been made capital of the new Grand Duchy and a newspaper in Turku no longer represented a larger unit, but rather was more locally oriented. This local orientation is also visible in another data-specific finding. One of the most common bigrams with the word "Finnish" stem from announcements regarding meetings of the Imperial Finnish Economic Society. Like in the case of Finnish-language masses in Stockholm in the 1770s, this repeated feature in the newspaper affects results in an unexpected way, especially as the Economic Society could reasonably be classified as geography, nation or state. Our classification according to geography is consequently visible in the geography column.

5. Conclusions

One difficulty in the historical discussion about how Finland and Finishness was understood from the eighteenth century to the nineteenth century, is that, depending on the point of view of the interpreter, different aspects of it could be emphasized. For historiography that sought to emphasize the existence (and often even greatness) of a Finnish past, it was important to underline the continuities in talking about Finnishness [32]. Those who challenged that historiography from the 1960s onwards, often emphasized the institutions that were established in the period after 1809 [29]. Our perspective that draws on historical language use to uncover historical understandings is not new in the sense that previous research has also paid attention to language [3, 21], but our method which is based on raw frequencies of bigrams and is transparent aides the analysis of where the continuities and discontinuities in the historical understanding are.

As anticipated, the immediate access to original occurrences proved to be a useful affordance. Many of the head nouns were such that they did not disambiguate between different senses of Finnish and required checking by close reading. Some of the contexts were ambiguous in themselves. The main source of uncertainty regarding the results stemmed from the relatively small size of datasets, which was exaggerated by the low quality of the OCR. However, we retain that our main observations would likely hold, even if more data were available, thanks to the fact that exact information about the sources of uncertainty was readily available. This meant that we were able to approximate their possible scope of influence. With improved OCR and a border selection of sources, it would be possible to refine the picture further.

Already now our comparisons show that the notion of Finnishness was differently weighed in Turku, Stockholm and Lund in the 1770s and that the conceptualization of Finnishness did expand after the so-called divorce of 1809. With regard to bigrams that relate to the state, we see an increase in variation within the category and, perhaps unintuitively, we see that Finnishness became less language-oriented in the 1830s data. The explanations relate to the new dynamic between the Swedish and Finnish languages in Finland after 1809, but as they are not straightforward, we believe a further comparison with other newspapers from the same period would clarify this shift further.

6. References

[1] Klinge, M.: Senaatintorin sanoma: Tutkielmia suuriruhtinaskunnan ajalta. Otava, Helsinki (1986).

[2] Topelius, Z.: Äger det finska folket en historie? Joukahainen. Ströskrift utgifven af Österbottniska Afdelningen. 2, 189–217 (1843).

[3] Marjanen, J.: Finland som begrepp och retorik. Historisk Tidskrift för Finland. 102, 533–544 (2017).

[4] Tahmasebi, N., Borin, L., Jatowt, A.: Survey of computational approaches to lexical semantic change detection, https://doi.org/10.5281/zenodo.5040302, (2021).

[5] Church, K., Gale, W., Hanks, P., Hindle, D.: Using statistics in lexical analysis. In: Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. pp. 115–164. Erlbaum (1991).

[6] Liu, D.: Is it a chief, main, major, primary, or principal concern? A corpus-based behavioral profile study of the near-synonyms, (2010). https://doi.org/10.1075/ijcl.15.1.03liu.

[7] Svenska Akademiens ordbok (SAOB).

[8] Geeraerts, D.: Vagueness's puzzles, polysemy's vagaries. Cognitive Linguistics. 4, (1993).

[9] Lakoff, G.: Women, Fire, and Dangerous Things. University of Chicago Press, Chicago (1987).

[10] Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., Mäkelä, E.: Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine. 22, (2016). https://doi.org/10.1045/july2016-paakkonen.

[11] Borin, L., Forsberg, M., Roxendal, J.: Korp – the corpus infrastructure of Språkbanken. In: Proceedings of LREC 2012. Istanbul: ELRA. pp. 474–478 (2012).

[12] Doucet, Antoine, Gasteiner, Martin, Granroth-Wilding, Mark, Kaiser, Max, Kaukonen, Minna, Labahn, Roger, Moreux, Jean-Philippe, Muehlberger, Guenter, Pfanzelter, Eva, Therenty, Marie-Eve, Toivonen, Hannu, Tolonen, Mikko: NewsEye: A digital investigator for historical newspapers. (2020). https://doi.org/10.5281/ZENODO.3895269.

[13] Tommila, P., Landgrén, L.-F., Leino-Kaukiainen, P.: Suomen lehdistön historia 1. Sanomalehdistön vaiheet vuoteen 1905. Kustannuskiila, Kuopio (1988).

[14] Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., Tolonen, M.: A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917. JEPS. 4, 54–77 (2019). https://doi.org/10.21825/jeps.v4i1.10483.

[15] Lundell, P.: Pressen i provinsen: Från medborgerliga samtal till modern opinionsbildning 1750–1850. Nordic Academic Press, Lund (2002).

[16] Kanner, A.: Meaning in distributions: A Study on Computational Methods in Lexical Semantics. University of Helsinki, Helsinki (2022).

[17] Katz, J.J., Fodor, J.A.: The Structure of a Semantic Theory. Language. 39, 170 (1963). https://doi.org/10.2307/411200.

[18] Langacker, R.W.: Concept, Image, and Symbol: The Cognitive Basis of Grammar. De Gruyter Mouton (2010). https://doi.org/doi:10.1515/9783110857733.

[19] Pustejovsky, J.: Type Theory and Lexical Decomposition. In: Pustejovsky, J., Bouillon, P., Isahara, H., Kanzaki, K., and Lee, C. (eds.) Advances in Generative Lexicon Theory. pp. 9–38. Springer Netherlands, Dordrecht (2013). https://doi.org/10.1007/978-94-007-5189-7_2.

[20] Nordin, J.: Ett fattigt men fritt folk: Nationell och politisk självbild i Sverige från stormaktstid till slutet av frihetstiden. B Östlings bokförl. Symposion, Stockholm Stehag (2000).

[21] Nordin, J.: Finland och riket. Formell och strukturell ojämlikhet. In: Engman, M. and Villstrand, N.E. (eds.) Maktens mosaik: enhet, särart och självbild i det svenska riket. pp. 211–231. Svenska litteratursällskapet i Finland, Helsingfors (2008).

[22] Marjanen, J.: Ekonomisk patriotism och civilsamhälle: Finska hushållningssällskapets politiska språkbruk i europeisk kontext 1720–1840. Finska Vetenskaps-Societeten, Helsinki (2022).

[23] Manninen, J.: Valistus ja kansallinen identiteetti. Aatehistoriallinen tutkimus 1700-Luvun Pohjolasta. Suomalaisen Kirjallisuuden Seura, Helsinki (2000).

[24] Marjanen, J.: Kuinka suomalainen oli Suomen ensimmäinen sanomalehti?, https://www.newseye.eu/fi/blogi/news/kuinka-suomalainen-oli-suomen-ensimmaeinen-sanomalehti/, last accessed 2022/11/02.

[25] Steinby, Torsten.: Finlands tidningspress: En historisk Översikt. Söderström, Helsingfors (1963).

[26] Screen, J.E.O.: The Army in Finland during the last decades of Swedish rule, 1770-1809: J. E. O. Screen. SKS / Finnish Literature Society, Helsinki (2007).

[27] Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., Nevalainen, T.: Wrangling with non-standard data. In: Reinsone, S., Skadiņa, I., Baklāne, A., and Daugavietis, J. (eds.) Proceedings of the Digital Humanities in the Nordic Countries 5th Conference. pp. 81–96. CEUR-WS.org, Germany (2020).

[28] Krusius-Ahrenberg, L.: Finland och den svensk–ryska allianspolitiken intill 1830/31 års polska revolution. Historiska och litteraturhistoriska studier. 21–22, 153–346 (1946).

[29] Jussila, O.: Maakunnasta valtioksi: Suomen valtion synty. WSOY, Porvoo (1987).

[30] Hemstad, R.: Scandinavianism. Mapping the Rise of a New Concept. Contributions to the History of Concepts. 13, 1–21 (2018). https://doi.org/10.3167/choc.2018.130102.

[31] Edgren, H.: Publicitet för medborgsmannavett: Det nationellt svenska i Stockholmstidningar 1810–1831. Acta Universitatis Upsaliensis, Universitetsbiblioteket [distributör], Uppsala (2005).

[32] Fewster, D.: Visions of past glory: Nationalism and the construction of early Finnish history. Finnish Literature Society, Helsinki (2006).