Digging Deeper into the Finnish Parliamentary Protocols – Using a Lexical Semantic Tagger for Studying Meaning Change of Everyman's Rights (allemansrätten)

Kimmo Kettunen[1] and Matti La Mela[2]

[1] The National Library of Finland, DH Research
[2] Semantic Computing Research Group, Aalto University
kimmo.kettunen@helsinki.fi,matti.lamela@aalto.fi

Abstract. This paper analyses the protocols of the Parliament of Finland 1907–2000. They have been digitised and published as open data by the Finnish Parliament in 2018[1]. In the analysis we use a novel tool, a semantic tagger for Finnish – FiST [1]. We describe the tagger generally and show results of seman-tic analysis both on the whole of the parliamentary corpus and on a small subset of data where everyman's rights (a widely used right of public access to nature) have been the main topic of parliamentary discussions. Our analysis contributes to the understanding of the development of this "tradition" of public access rights, and is also the first study utilizing the Finnish semantic tagger as a tool for content analysis in digital humanities research. Keyword search shows first that that the discussion of everyman's rights has had three different peak peri-ods in the Finnish Parliament: 1946, 1973, and 1992. Secondly, the contents of the discussions have different nature for all the periods, which could be clearly detected with FiST and keyness analysis.

Keywords: Parliamentary Proceedings, Everyman's Rights, Allemansrätten, Semantic Tagging, Parliament of Finland.

1 Introduction

Language technology has been used for semantic analysis of languages during the last few decades quite a lot, but still proper tools for semantic or content analysis of less resourced languages may be scarce. This, for example, is the situation for the Finnish language: content analysis tools or more generally semantic resources for Finnish are rare [2].[2] In this study we use a novel prototype lexical semantic tagger, FiST [1], for analysis of parliamentary protocols 1907–2000 of the Parliament of Finland. Firstly we analyse, whether the quality of the OCRed documents is good enough for large scale content analysis using the semantic tagger. Secondly, we will perform initial analysis of the contents of the protocols and try to establish, whether lexical coverage

[1] https://avoindata.eduskunta.fi/#/fi/digitoidut/

[2] Kettunen [1] lists most of these tools for research, and we do not discuss them here.

of results of the semantic tagger is high enough in the data to be useful for detailed analysis. Thirdly, we will analyse a particular subset of the parliamentary protocols to study the changes in the syntagmatic semantic neighborhood [3] or distribution [4] of "everyman's rights" (allemansrätten), a public access right to nature [5–6]. The everyman's rights are commonly understood as a legal-cultural tradition in the Nordic countries [cf. e.g. 6–9] yet, we know only little about the history of this right [cf. 10: 215–218]. We contribute to the scholarship by studying the changes in the public uses of the term in Finland in the twentieth century, and in this way, unfold the narrative of a traditional right. Our analysis of everyman's rights is the first study utilizing the Finnish semantic tagger as a tool for content analysis in digital humanities research, and we aim to see how useful FiST's semantic scheme is for this purpose.

Digital versions of national and multinational parliamentary texts have been compiled and analysed at least since the publication of the first multilingual EuroParl version [11]. Parliamentary texts have been analysed from different viewpoints, which include e.g. topic detection, metadata enhancement, sentiment analysis, reputation defence, gender studies, textometrics, political stance and political group differences [12–14]. Many times the work seems to be concerned with creation of suitable annotation and metadata for the data [12]. Content wise annotated parliamentary corpora are still rare [15] and analysis of the data is many times based on current statistically oriented NLP approaches.

Our tools and methods in this study owe much to traditional corpus analysis. We have available a lexical semantic tagger of Finnish, and we use it for production of annotated data out of the raw parliamentary proceedings of the Finnish Parliament. The output of the semantic annotation is analysed both with corpus statistics and intellectually, i.e. both with distant and close reading, as the parlance in digital humanities goes. We assume that the semantic annotation of our tagger provides a better way for analysis of the data than e.g. general machine learning tools such as Mallet[3], which introduce topics as un-interpreted and ungrouped keywords taken from the actual texts[4]. Semantic labelling offers us a possibility to generalise the findings easily to meaningful categories, even if the semantic categorisation used is a general linguistic model that has not been tailored particularly to any specific use. From experience of the use of the semantic categorization in English, however, we know that the categorization has been useful in many kinds of studies[5].

Our hypothesis in this study is that we manage to give sense to the parliamentary debates through semantic tagging with FiST. More specifically, the semantic tagging and keyness analysis allows us to describe the shifts in the meaning of everyman's rights. Today, everyman's rights are part of the identity, way of living and national brand-building in the Nordic countries [18–19]. It refers to the legal principle, accord-

[3] http://mallet.cs.umass.edu/

[4] The same applies also for one of the most used unsupervised machine learning topic detection methods, Latent Dirichlet Analysis (LDA, [16–17]). The bag(s) of words given as topics by the topic modelling software can be grouped under common themes or titles by a group of evaluators afterwards, but that is not without its problems. Results of LDA are also partly dependent on choice of parameters and their tuning.

[5] Cf. publication listing at http://ucrel.lancs.ac.uk/wmatrix/#apps

ing to which everyone is allowed to roam in nature and take use of wild natural resources even without the consent of the landowner [5]. Even though narrated as an age-old institution, we know that the legal concept itself developed only from the 1930s onwards – with the on-going urbanization, growth of free time and new practices of access to nature. [8, 19]. We expect to see in our period the stabilization of a core for the everyman's rights, but also the broadening of the range of use of the term in public talk.

2 FiST – a Prototype Semantic Tagger of Modern Standard Finnish

Kettunen [1] has implemented a prototype lexical semantic tagger for Finnish, FiST. The tagger has been developed using freely available components: FinnPos morphological tagger [20] and a 46K Finnish semantic lexicon published in 2016 [21–22]. The Finnish semantic lexicon has been developed using the lexicon of the English Semantic Tagger (The EST) of University of Lancaster as a model. This semantic tagger was developed at the University Centre for Corpus Research on Language (UCREL) at Lancaster University as part of the UCREL Semantic Analysis System (USAS[6]) framework. The semantic lexicon of the USAS framework is based on the modified and enriched categories of the Longman Lexicon of Contemporary English [23].

The implementation of FiST uses Omorfi and FinnPos for morphological analysis of Finnish words. After the morphological analysis phase words from the 46K semantic lexicon are matched against the morphologically unambiguous base forms. FiST is a first version of the semantic tagger and it lacks still some features, especially word sense disambiguation [24] and proper handling of compounds. However, it achieves already a lexical coverage of 82–91% in several types of texts of modern standard Finnish [1].

Semantic tagging of FiST is based on the idea of semantic (lexical) fields. Wilson and Thomas [25: 54] define a semantic field as "a theoretical construct which groups together words that are related by virtue of their being connected – at some level of generality – with the same mental concept". According to Dullieva [26] "a semantic field is a group of words, which are united according to a common basic semantic component", cf. also [27]. Semantic lexicon of USAS is divided in to 232 meaning classes or categories, which belong to 21 upper level fields. Table 1 shows one upper level semantic field, Money & Commerce, and its meaning classes[7]. Alphanumeric abbreviations in front of the meaning classes are the actual hierarchical semantic tags used in the lexicon. According to Piao et al. [28], the depth of the semantic hierarchical structure is limited to a maximum of three layers, since this has been found to be the most feasible approach. The major 21 discourse domains used in the USAS are listed in Table 2.

[6] http://ucrel.lancs.ac.uk/usas/

[7] USAS Semantic Tag Set, http://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf

Table 1. Semantic field of Money & Commerce in the USAS semantic lexicon.

I MONEY & COMMERCE	I2.1 Business: Generally
I1 Money generally	I2.2 Business: Selling
I1.1 Money: Affluence	I3 Work and employment
I1.2 Money: Debts	I3.1 Work and employment: Generally
I1.3 Money: Price	I3.2 Work and employment: Professionalism
I2 Business	I4 Industry

Table 2. Top level domains of the USAS tag set.

A	General & Abstract Terms	N	Numbers & Measurement
B	The Body & the Individual	O	Substances, Materials, Objects & Equipment
C	Arts & Crafts	P	Education
E	Emotional Actions, States & Processes General	Q	Linguistic Actions, States & Processes
F	Food & Farming	S	Social Actions, States & Processes
G	Government & the Public Domain	T	Time
H	Architecture & Building, Houses & the Home	W	The World & Our Environment
I	Money & Commerce	X	Psychological Actions, States & Processes
K	Entertainment & Sports and Games	Y	Science & Technology
L	Life & Living Things	Z	Names & Grammatical Words
M	Movement, Location, Travel & Transport		

The parliamentary protocols of the Finnish Parliament 1907–2000 have been digitised and published as open data by the Parliament of Finland in June 2018[8]. The documents record the work of the unicameral Parliament of Finland established in 1906, and offer a unique view on the key national legislative reforms and public issues of the twentieth century. The documents contain the law proposals and petitions, preparatory and committee work, and the transcribed plenary debates. The digitisation has been produced in the Parliament. As a paper collection, the data consists of about 1.9 million pages and 1830 bound books, as digitised data about 91.4 GB, in Finnish and Swedish. The Finnish data has ca. 479 million tokens, punctuation included. The data is available both for search and downloading on the open data pages of the Finnish Parliament. The protocols of 1907–1975 are divided into three categories: minutes, documents and appendices. Years from 1975 onwards conform to a different system:

[8] http://avoindata.eduskunta.fi/digitoidut/

documents were referred to with capital letters A–F with numbers showing possible subparts. [29]. In general, the Finnish parliamentary data contains currently no structured information about the speakers, parties or subject areas, which would greatly benefit all research done with the parliamentary proceedings.

The Web pages of the Parliament contain only pdf versions of the digitised data. Text versions of the data in our study have been produced using pdftotext utility[9].

3.1 Semantic Tagging of the Parliamentary Data

Kettunen [1, 30] has shown that FiST is capable of analysing robustly different types of written Finnish text regardless of their genre. These papers analysed newspaper texts, web discussion forum texts, Bible translation, fiction, and proceedings of the European Parliament, among others, and the amount of data ranged from ca. 6000 words to 45 million words. Part of the data was supposedly outside the scope of the semantic lexicon of the tagger, but anyhow the lexical coverage was most of the time over 80% and many times 90–91%. Worth noticing is that the proceedings of the European Parliament[10] with its 28.6 million words got coverage of 90.9%. This particular example data implies clearly, that parliamentary proceedings should be analysable with FiST.

La Mela [31] has earlier analysed the Finnish parliamentary proceedings with morphological analysers. The results have been produced with LAS[11], a Linguistic Analysis Command-Line Tool that wraps up several existing linguistic analysis tools in a single package. La Mela's analysis showed that the word level recognition of the data is relatively high most of the time. Between years 1918 and 1927 there is a clear drop in recognition, but otherwise the recognition rate is between 70 and 90 per cent most of the time and for the few last years clearly over 90%. Only Finnish words were recognised, which lowers the recognition level slightly, while the MPs used also Swedish in the debates [31]. The recognition rate can be considered a rough estimation of the quality of the digitization [32]. For clean modern standard Finnish the recognition rate could be around 95% [33].

With this background of FiST's lexical coverage in general and morphological analysis of the parliamentary protocols, possibilities for semantic tagging of the protocols looked reasonably promising and we started to annotate them with FiST. As a result of the tagging, words of the texts are either given semantic tag annotation or tag Z99 as a mark of an unknown word for the semantic lexicon. A short example of the tagger's output is shown in Table 3.

[9] http://www.xpdfreader.com/

[10] http://www.statmt.org/europarl/archives.html#v6

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

Table 3. An example of FiST's analysis.

Original running text		Results of FiST, words in base form
1.	Suomen	suomi Noun Z2/Q3
2.	eduskunnan	eduskunta Noun G1.1
3.	vastaus	vastaus Noun Q2.1
4.	.	. PUNCT
5.	Hallituksen	hallitus Noun G1.1/S5+ S7.1+/S5+
6.	esitys	esitys Noun Q2.2 X7+ K4 X9.2

We can see different types of tags in the result. The second and the third word token have unambiguous single tags; the others have multiple tags, as semantic ambiguity is not resolved in the tagger. In most of the cases the first tag is probably the right one, as the most frequent tag for each word is the first one in the Finnish semantic lexicon [21: 74]. In the literature of word sense disambiguation, this is known as the most frequent meaning baseline, which is many times hard to outperform with disambiguation methods [24].

While running the parliamentary documents through FiST we noticed that the figures given by morphological recognition of LAS may be too optimistic. Figure 1 shows lexical coverage of the semantic tagger with all the main protocols (minutes) of the Parliament 1907–2000 and with the whole data.

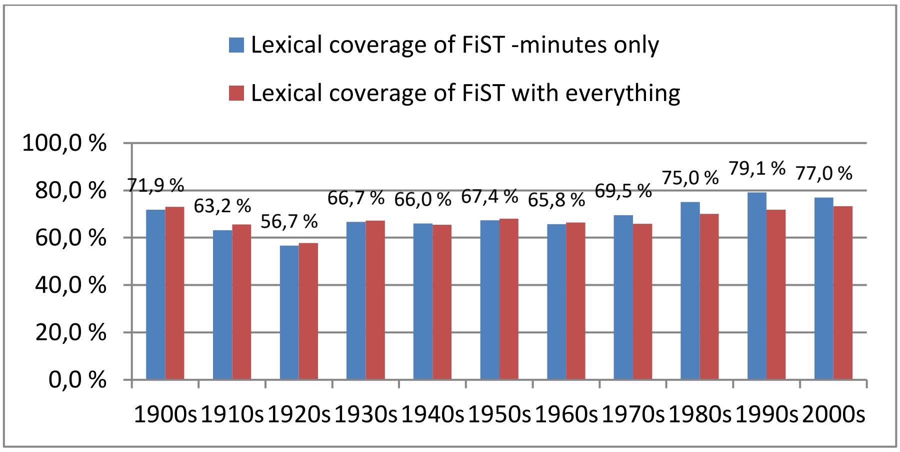


Fig. 1. Lexical coverage of FiST with the parliamentary data: percentages shown for the minutes.

Figure 1 shows that lexical coverage of the semantic tagger varies from ca. 57 to 79 per cent[12]. Earlier minutes of the Parliament have overall a slightly worse recognition

[12] Lexical coverage counting with FiST differs from morphological figures of La Mela [31], as we have here omitted punctuation from the overall count. Thus the overall figures are clearly lower already for this reason.

rate than the whole data, but from 1970s onwards minutes have a clearly better recognition rate. Most of the time the differences are small. Average recognition rate for the minutes is 68.9%, and for the whole data 67.7%.

Based on browsing of the list of unknown words in the results, the main reasons for differences in the recognition rates of morphological analyser and semantic tagger are these word types that cannot be recognised by FiST:

- Broken words in the data due to line ending hyphenation – morphological analysis probably recognises part of the errors as "words".
- OCR errors in the data – morphological analysis probably recognises part of the errors as "words".
- Compound words: many of these could not be recognised with FiST, because FinnPos did not split compounds to their constituent parts when analysis runs were performed, and the lexicon of FiST contains only part of possible compounds; morphological analyser has a higher coverage with compounds.
- Swedish words: there is Swedish in the data even if the main language of the data is Finnish.
- Names of persons (e.g. last names of the members of the parliament) that are not recognised are mentioned often and repeatedly – morphological recogniser has a larger lexicon for these.
- Abbreviations in the texts, e.g. *n:o*, 'number', is very frequent; many of the frequent abbreviations are spelled differently in today's Finnish.

Part of these errors – especially hyphenation and abbreviations – could be corrected with preprocessing, but we have not tried to improve quality of the data in any way. The number of errors may have an impact on the results of our analysis, but we believe that the achievable level of semantic tagging makes the study feasible. Moreover, as described below, we will process our data and use context windows of Finnish text, which will improve the recognition rate in our study corpus.

3.2 Everyman's Rights in the Data

In the rest of this article, we employ FiST for analysing the use of the term everyman's rights in the parliament data. This is done by studying the semantic tagging in context windows around the token "everyman" in the data.

We first searched for all the occurrences of everyman in the minutes of the plenary debates. The keyword used in the search was "*okamie*", which is the truncated keyword for different inflections of Finnish everyman.[13] The keyword appeared 1003 times in the minutes of the Parliament, and resulted into 5 false hits, which were re-

[13] The written form everyman's rights ("jokamies" written together), instead of "every man's rights" ("joka miehen"), became commonly used for the concept during the latter half of the twentieth century [31].

moved from the results.[14] Figure 2 shows the number of hits in the timeline of the collection (since the first hit in 1932).

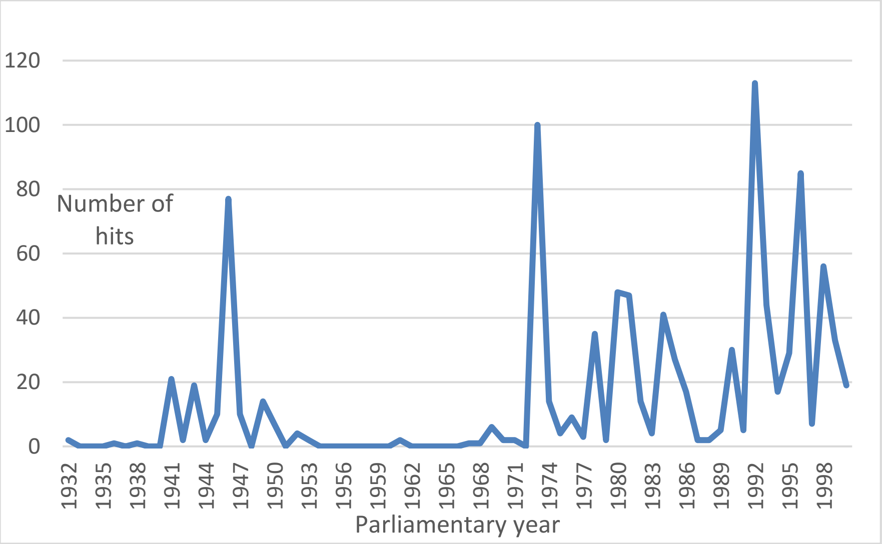


Fig. 2. Search hits per parliamentary year for everyman: "*okamie*" (1932–2000).

The timeline confirms previous observations that the term everyman's rights itself became commonly used only during the second half of the twentieth century [18]. The peaks in the figure depict the nature of parliamentary work, and show how the term has been more topical during certain years and when specific legislative projects were debated in the Parliament. The results show three clear peaks in discussing the everyman's rights: 1946, 1973, and 1992. Also years 1996 and 1998 are among the most notable years in search hits. Based on this, we focused our analysis on three decades, 1940s, 1970s, and 1990s, and the specific peak years in the debates: 1946, 1973, and 1992.

Next, we formed our *everyman corpus*, which we use in the analysis below. The corpus contains 40 120 words. The corpus consists of context windows, which capture 20 words surrounding the search hits ("*okamie*") on both left and right side. We used only Finnish minutes of the plenary debates, and preprocessed the word tokens in the documents by removing other characters than letters or hyphens (when used for compound words), and tokens of only one character. Then we processed the corpus with FiST. If there were several FiST tags for one token (ambiguity), we picked only the first tag relying on the principle of the most frequent sense of the lexicon. If the first tag contained two discourse domains (separated with slash /), we stored both and gave them weight of 0.5 in the analysis.

[14] These were OCR errors (joka"mietinnön → jokamietinnön) or search string errors ("pokamies", bachelor).

The lexical coverage in the everyman corpus follows the trend presented above in Figure 1, but it is somewhat higher than in the minutes of the parliament in total, being 86.1% for the complete everyman corpus. This is due to the preprocessing of the tokens and the capturing of tokens in Finnish language debates (with less technical or procedural vocabulary present).

Finally, we built three comparative corpora for our analysis in section 4.2. These comparative corpora consist of the complete minutes from three years preceding and succeeding the peak years we analyse: the years between 1943 and 1949 (for the year 1946), 1970–1976 (for 1973), and 1989–1995 (for 1992). In total, the corpora consist of 9.2 M, 17.6 M, and 18 M word tokens, respectively. Lexical coverages of FiST for these three corpora are 65.9, 69.2 and 79 per cent, respectively.

4 Semantic Analysis of Everyman's Rights through Parliamentary Debates

In this section, we focus on the development of the Finnish concept of everyman's rights in the twentieth century, and use FiST to study the changes in its semantic neighborhood (syntagmatic relations [3]; or distributions, in the vector space parlance, [4]). We are interested in whether we are able to describe the content of everyman's rights, detect in which legislative debates the term appears and to portray changes in the meaning of the term when used in the Parliament.

4.1 Methods Used in Analysis

We have already described our analysis methods to some extent, but for clarity we collect and explain our main methods of text analysis in this section.

We have so far explained the basics of our main content analysis tool, a prototype lexical semantic tagger of Finnish. It produces semantically categorised words in base form, and the semantic categorizations give a better way to inspect themes of discussion than plain words without any markings [34–35]. The semantically tagged output of the parliamentary proceedings could be studied in many ways; these are the ways we use[15].

We have four different corpora for the analyses as explained in section 3.2. The small hand compiled *everyman corpus* is our study corpus and the three different comparative corpora are the reference corpora in the corpus analysis parlance [36].

Meaning of the concept *jokamiehen (oikeus)* is studied in a syntagmatic way [3]. This means that the semantically interpreted word surroundings (contexts) of the token *jokamies* ("everyman") are analysed. This way we get a distributional view of the co-occurences of semantically categorised words in the contexts of *jokamies*.

For obtaining the distributions, we need to define a meaningful context for studying. As our context window we have chosen 20 words on the left and right side of

[15] As was seen in Table 1, current output of FiST consists of the lemma or base form of the input word, its word class and the semantic marking. If needed, we could also use more linguistic marking from the output of FinnPos.

occurrences of token *jokamies*. From a linguistic point of view we can justify the choice of our word window size in the *everyman corpus* as follows:

- Quality of the data is not good, and sentence boundaries cannot be found automatically many times, thus we need to rely on mere word counting for the context.
- Most interesting things related to the notion of *jokamies* are probably said near to the word itself, either immediately before or after it.
- The window size is a 40-gram, which is 8 to 20 times larger than word n-grams used many times in NLP (the size of n varying usually from 2 to 5, cf. e.g. [37–39] – a large enough context is needed for meaningful syntagmatic relations.
- According to sentence length statistics of Niemikorpi [40: 176–177] Finnish of different genres (46 text types from 1960s) have mean average sentence length of 12.83 words. The shortest mean length of sentences in Niemikorpi's data is 5.8 words, longest 24.3. Thus we can assume that our window length covers at least about one sentence before and after token *jokamies*. This can be considered a meaningful contextual window. A very large context window, e.g. the whole document, would be futile, because in a large context nearly every word can co-occur with every other; and on the other hand, a very narrow context, like few words, would cause a serious sparse data problem, while words co-occur very rarely with each other in a small textual window [3].
- However, it is possible, that a larger context window, say 50 words on both sides of *jokamies*, could be informative, too, but we leave this for a later stage.

The three reference corpora are selected for three years preceding and succeeding the years we analyse, while we wanted the comparative corpus to surpass the context of the legislative process where everyman's rights were discussed, but to limit our comparison to the specific public language use of the time. After 1930, the Parliaments in Finland have usually functioned for 3 or 4 years (the complete term is four years).

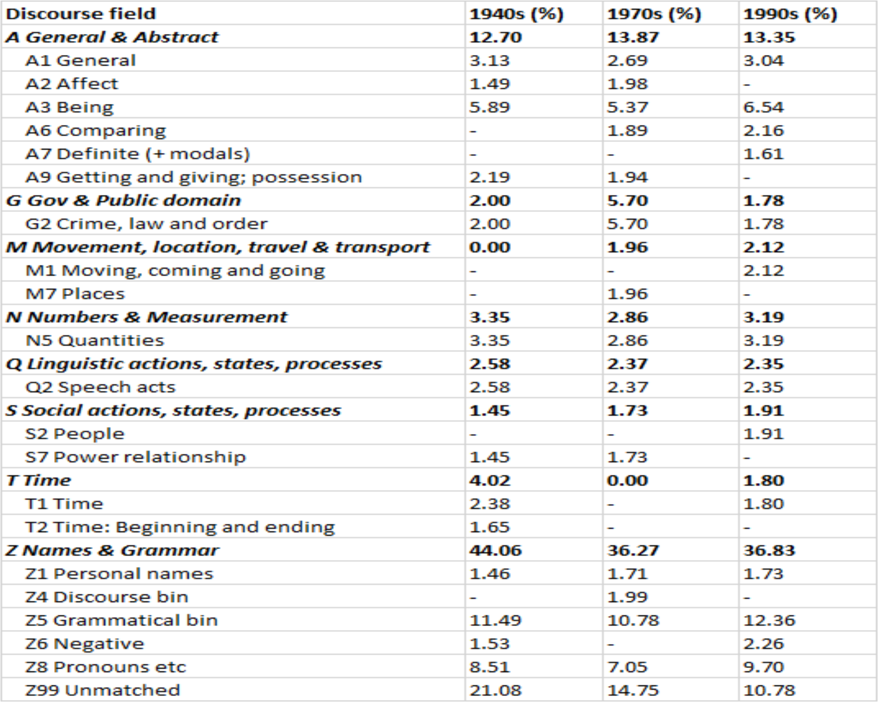
In finding of the most important words related to *jokamies* we use keyness analysis, which concerns statistically significant differences or similarities in the relative occurrences of words in two corpora [36]. The keyness method in our study is extended from keywords (plain linguistically unmarked word tokens in the text) to key semantic domains marked by FiST [35], which gives us a better insight into the contents of the texts: the overall trends are easier to analyse.

The output of the semantic annotation is analysed both with corpus statistics and intellectually, i.e. both with distant and close reading in a digital humanities manner. Our methodology and its reasoning resemble ideas that e.g. de Bolla et al. [42] discuss and which they consider to follow good practice in digital humanities: tool construction (or choice of tools, too) is driven in the first instance by research questions.

4.2 FiST Tagged Context Windows Per Decade

We studied first the occurrences of the FiST tagged categories per decade in the everyman corpus. This will give us a preliminary idea of the semantic content of the corpus and gives first insights about the similarities and differences between the decades. Table 4 portrays the shares (discourse field tags per total tags) for the top 15 categories for 1940s, 1970s, and 1990s. For the level of analysis, we selected the main discourse field (A-Z) and the first subdivision.

Table 4. The shares of top 15 first-division subfields tagged by FiST in the everyman corpus for 1940s, 1970s and 1990s.



Discourse field	1940s (%)	1970s (%)	1990s (%)
A General & Abstract	**12.70**	**13.87**	**13.35**
A1 General	3.13	2.69	3.04
A2 Affect	1.49	1.98	-
A3 Being	5.89	5.37	6.54
A6 Comparing	-	1.89	2.16
A7 Definite (+ modals)	-	-	1.61
A9 Getting and giving; possession	2.19	1.94	-
G Gov & Public domain	**2.00**	**5.70**	**1.78**
G2 Crime, law and order	2.00	5.70	1.78
M Movement, location, travel & transport	**0.00**	**1.96**	**2.12**
M1 Moving, coming and going	-	-	2.12
M7 Places	-	1.96	-
N Numbers & Measurement	**3.35**	**2.86**	**3.19**
N5 Quantities	3.35	2.86	3.19
Q Linguistic actions, states, processes	**2.58**	**2.37**	**2.35**
Q2 Speech acts	2.58	2.37	2.35
S Social actions, states, processes	**1.45**	**1.73**	**1.91**
S2 People	-	-	1.91
S7 Power relationship	1.45	1.73	-
T Time	**4.02**	**0.00**	**1.80**
T1 Time	2.38	-	1.80
T2 Time: Beginning and ending	1.65	-	-
Z Names & Grammar	**44.06**	**36.27**	**36.83**
Z1 Personal names	1.46	1.71	1.73
Z4 Discourse bin	-	1.99	-
Z5 Grammatical bin	11.49	10.78	12.36
Z6 Negative	1.53	-	2.26
Z8 Pronouns etc	8.51	7.05	9.70
Z99 Unmatched	21.08	14.75	10.78

In general, we did not see major differences between the general discourse fields in different decades. The “A General” and “Z Names & Grammar” categories are the largest groups. The share of the latter decreases when we move to the 1990s: this is mainly related to the better recognition quality of the tagged words. We can detect some movement inside the top tagged categories per decade. First, we see that the category A9 referring to possession is present only among the top categories of the first two decades. Second, the “M Movement, location, travel” category is present in the latter decades, in the 1970s and 1990s.

When we look at the unique categories of the decades, there are small nuances. Close reading of the results enables us to see what tokens have been included in the categories.

The everyman of the 1940s was related to time, periods of time (T1, T2), and possession (A9). Among the tagged tokens, we found "to continue/continuation" (*jatkaa, jatkuminen*), "still/further" (*vielä, edelleen*) among the common T2 temporal terms, and "to receive/to acquire" (*saada, hankkia, ansaita)*, "proprietor" (*omistaja*), "to own" (*omistaa*) for the A9 possession category tokens.

In the 1970s, the use of "everyman" seems to shift towards its legal definition, and more typical areas of everyman's rights such as location and movement. We see a particularly large share of tokens classified as "G2 Crime, law and order". The movement category is related to the static category of "M7 Places", whereas in the 1990s, movement is about "M1 Moving, coming and going". We find the tagged tokens "right" (*oikeus*), "law" (*laki*), in the G2 law and order category, and "state" (*valtio*), "(camping) area" (*retkeily/alue*), and also "Nordic country" (*Pohjoismaa*) for the M7 movement category.

Finally, in the 1990s, we find the category "S2 People", which is not among the main tagged categories of the previous decades. The S2 people category of the 1990s include tokens such as "Finnish" (*suomalainen*), "(Central-) European" (*keski-/eurooppalainen*), "outsider" (*ulkopuolinen*), "citizen" (*kansalainen*), "human being" (*ihminen)*. These refer to mechanisms of grouping and exclusion between different groups.

The shares of tagged tokens around the term "everyman" remain stable from decade to decade. Also, the unmatched category Z99 grows smaller, which may affect the FiST categories unequally. The small differences, however, hint at changes in the uses of "everyman" in the parliamentary and public discussion.

4.3 FiST Tagged Context Windows in 1946, 1973, and 1992

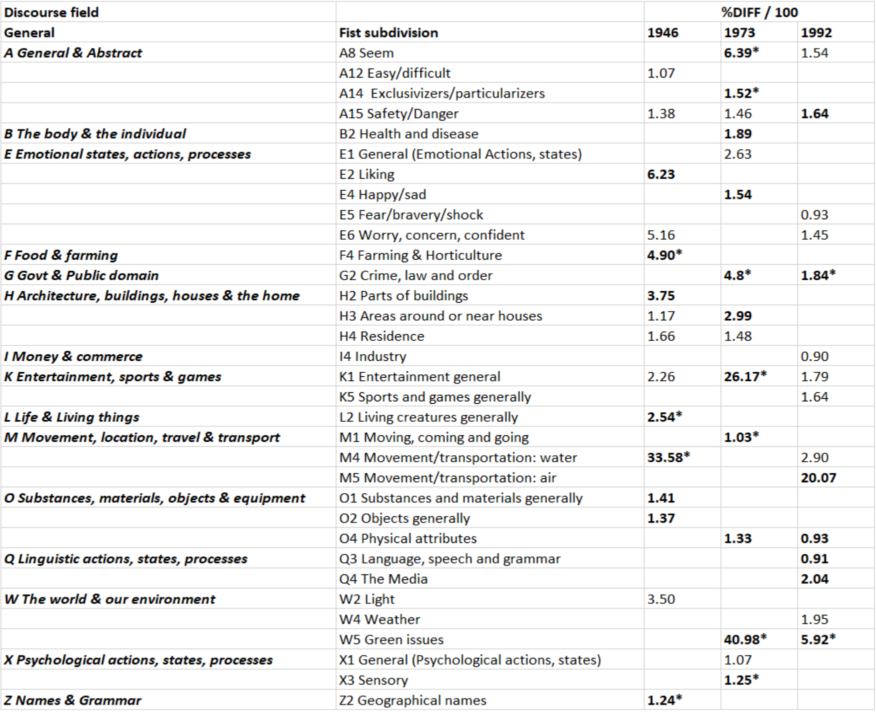
After inspecting the shares of the top discourse fields per decade, we focused on yearly results in the everyman corpus, and the legislative context where the term was used. For more nuanced results, we studied the relative occurrence of the FiST tags from the three years with most search hits: 1946, 1973, and 1992. We used keyness analysis, which is about presenting statistically significant differences or similarities in the relative occurrences of words in two corpora [34]. We compared the yearly results of the everyman corpus to the corresponding comparative corpus (the complete minutes from three years preceding and succeeding our year of study). In our case, we were interested in the categories that were more frequent in the everyman corpus than in the minutes (comparative corpus). Table 5 presents the 15 most overrepresented categories in the everyman corpus based on %DIFF metric[16], and signals statistical

[16] The %DIFF value signals the normalised frequency of a category in the everyman corpus in comparison to the normalised frequency in the comparative corpus (complete minutes of +/-3 years): value 1 is twice the frequency of a category in the everyman corpus, value 2 three times the frequency. [36] For example, the category "F4 Farming & Horticulture" is in 1946

significance for the values. The non-significant categories are also listed, while they may open new questions for the analysis. These categories have low frequencies in the everyman corpus.

We noted that certain FiST categories are common for the everyman corpus in all three years. The general fields concerning Emotional discourse (E), Entertainment (K), Movement (M), Substances (O), and the World & environment (W) are generally common to everyman discourse, and can be understood as forming the core of the meaning of the term. These categories are, however, not equally represented, and they make the small differences we found in the previous section more visible. As we will see, the differences seem to reflect mainly the themes of the specific debates of the parliamentary year.

Table 5. The 15 most overrepresented subfields in the everyman corpus in relation to the complete minutes (+/- 3 years), for years 1946, 1973, 1992. Note: The metric used is %DIFF / 100. Asterisk (*) marks values with Bayes Factor of at least 2, and bolded values have p-value lower than 0.01.



Discourse field			**%DIFF / 100**	
General	**Fist subdivision**	**1946**	**1973**	**1992**
A General & Abstract	A8 Seem		**6.39***	1.54
	A12 Easy/difficult	1.07		
	A14 Exclusivizers/particularizers		**1.52***	
	A15 Safety/Danger	1.38	1.46	**1.64**
B The body & the individual	B2 Health and disease		**1.89**	
E Emotional states, actions, processes	E1 General (Emotional Actions, states)		2.63	
	E2 Liking	**6.23**		
	E4 Happy/sad		**1.54**	
	E5 Fear/bravery/shock			0.93
	E6 Worry, concern, confident	5.16		1.45
F Food & farming	F4 Farming & Horticulture	**4.90***		
G Govt & Public domain	G2 Crime, law and order		**4.8***	**1.84***
H Architecture, buildings, houses & the home	H2 Parts of buildings	**3.75**		
	H3 Areas around or near houses	1.17	**2.99**	
	H4 Residence	1.66	1.48	
I Money & commerce	I4 Industry			0.90
K Entertainment, sports & games	K1 Entertainment general	2.26	**26.17***	1.79
	K5 Sports and games generally			1.64
L Life & Living things	L2 Living creatures generally	**2.54***		
M Movement, location, travel & transport	M1 Moving, coming and going		**1.03***	
	M4 Movement/transportation: water	**33.58***		2.90
	M5 Movement/transportation: air			**20.07**
O Substances, materials, objects & equipment	O1 Substances and materials generally	**1.41**		
	O2 Objects generally	**1.37**		
	O4 Physical attributes		**1.33**	**0.93**
Q Linguistic actions, states, processes	Q3 Language, speech and grammar			**0.91**
	Q4 The Media			**2.04**
W The world & our environment	W2 Light	3.50		
	W4 Weather			1.95
	W5 Green issues		**40.98***	**5.92***
X Psychological actions, states, processes	X1 General (Psychological actions, states)		1.07	
	X3 Sensory		**1.25***	
Z Names & Grammar	Z2 Geographical names	**1.24***		

For 1946, the categories not appearing in the other years are “F4 Farming & Horticulture” and “L2 Living creatures generally”. The most distinct category is “M4 Movement-transportation water”. Moreover, we see location and material world expressed

almost six times (4.90) more common in the everyman corpus than in the minutes in general.

in the categories “H Architecture, buildings, houses & the home”, “Z2 Geographical names”, and Substance and Objects (O1 and O2). This is not surprising, as the debates in the early 1940s, and in 1946 in particular, regarded the principle of “Everyman’s fishing (right)” (*jokamiehen kalastus/oikeus*). These were temporary fishing rights created in the war years to alleviate food shortages, and they allowed households and war evacuees (the “Everyman fishers”) to fish without rights to the local fishing waters [43]. In 1946, the debate concerned whether the rights should be continued, made fixed, or rather to be ended.

It is notable, that the 1940s appears as the decade, when the term “everyman” became used more commonly in relation to nature and access to its resources [31]. We find the first appearance of “everyman” in the minutes in 1932. In the early century, the term “jokamies”, everyman, has been used to refer to the common man as in technical or law manuals for the everyman [18]. In a similar way, in the four search hits from the 1930s, the MPs alluded to everyman rather as synonym to common or ordinary people. In one case, “everyman” appeared as part of Hunting Law debate in 1932 about making hunting on state land possible for local people, and in a particular expression about “lands for hunting by the everyman” (“jokamiehen metsästettävät alueet”). The term, however, did not refer to universal rights, but to the possibilities of non-owning groups to hunt and use the resources. This is the meaning, which we found among the search hits of “everyman” in the 1940s.

When we move to the year 1973, the idea of roaming in the nature becomes very visible: the category “W5 Green issues” receives a major share, and also, the results in 1973 appear statistically most significant. As in 1943, the location and substances/objects categories are represented, but we find also the “K1 Entertainment” and “G2 Crime, law and order” categories. The fishing elements disappear, which is related to the ending of the particular fishing rights legislation of the 1940s. In 1973, the legislative debate (where the term everyman was used) concerned the enactment of the Outdoor Recreation Act. Close reading of the tagged tokens portrays this well: “camping” (*leirintä/alue*) and “outdoor recreation” (*ulkoilu*) are found in the category K1, “nature” (*luonto*) and “nature conservation” (*luonnonsuojelu*) in the category W5. Notably, the growth in law and order category G2 seems to regard the written form of everyman’s rights. In all results but one, the term is not a compound word, but written separately as “jokamiehen oikeus”, the right of the everyman, which makes the tagger find the token “right”.

The results from 1992 have the same key components as in 1943, the categories K entertainment, M movement, W5 Green issues. Also, the law and order category G2 is present, but this is not due to the separately written form. The majority of our search hits in 1992 are “everyman’s rights” written together (*jokamiehenoikeus*), which implies that the word “everyman” is not used independently anymore, but appears mainly in language-use as part of the expression everyman’s rights. We find the broadening of the entertainment category, which now includes the “K5 Sports and games”, and the appearance of the categories “Q4 The Media”. In close reading, we find tokens “hunting” (*metsästys*), “riding” (*ratsastus*), and “recreational fishing” (*virkistyskalastus*) in the category K5. The category Q4 is partly erroneous. For example, it tags the token “publication” (*julkaisu*), which refers to official publications not to journalistic work. Also, the word “Demari” classified as Q4, which is a colloquial term for a social-democrat but also the party newspaper’s name.

The debates in 1992 were about the hunting legislation, Finland’s EC membership, and property and company ownership by foreigners. These are not directly related to nature use, and portray how the term “everyman’s rights” has been used in debates regarding the country’s international status and citizens’ rights. This was visible in the previous section, where we examined the changes in the shares of the general discourse level, but we note also the declining role of the concrete categories such as location and nature. As the principle of everyman’s right was now commonly known and expressed with the term “*jokamiehenoikeus*”, it could be used as a more general rhetorical figure referring to the national tradition of access rights. Moreover, the 1992 results are statistically less significant and less different from the complete minutes than the results of the earlier years. This could imply that the term was used in the debates in a more abstract and more flexible way than in the previous decades.

5 Conclusion

In this paper we have used FiST – a lexical semantic tagger for Finnish – for analysing proceedings of the Parliament of Finland, which have been digitised for the years 1907–2000. At the same time, we were able to evaluate the quality of the digitised parliamentary documents, and found that the lexical coverage drops especially between 1918 and 1927, which results mainly from the weaker OCR quality. FiST reaches a lexical coverage of 57 to 79 percent in the complete material, which, although slightly low, we consider feasible for the purposes of our study.

Furthermore, we used FiST for studying the semantics of everyman’s rights, a Nordic principle of public access rights to nature. We formed a sub corpus, which contained +/-20 word windows around the Finnish term *jokamies*, everyman, and which we tagged with FiST. We focused on the decades and years, where we found most discussion taking place in the Parliament: 1940s (1946), 1970s (1973), and 1990s (1992). The FiST results were approached in two ways: by looking at the changes in the share and relative occurrence (keyness) of the FiST categories, and by close reading the results.

We found that during our period “everyman” became associated with the expression “everyman’s rights”. Already in the 1940s, the term referred to access rights and concrete outdoor environments, and was used to describe temporary fishing rights for locals. In the 1970s, the categories regarding nature, concrete outdoor environments and movement became even more central. The separate written form (*jokamiehen oikeus*, the right of everyman) generated excess recognition results about “rights”. In the 1990s, the compound form *jokamiehenoikeus* became the norm. The concept moved beyond the domestic access to nature discourse, and the term appeared in debates about national culture and institutions in the context of Finland’s EC membership. We found less nuanced and more varied categories, a sort of dilution with the complete minutes. We can say, thus, that the core meaning of everyman’s rights concerning access to nature was shared in the public debate at least in the 1970s, and incorporated elements related to national identity in the 1990s. Moreover, these results suggest that everyman's rights seem to have embodied the meaning of "public access to nature" later than commonly thought. This opens up new questions about how the political parties concretely shaped and defined the concept, which, however, would require annotation of the parliamentary data by speaker and party.

We used FiST for studying changes in the meaning of everyman's right. We were able to capture features about the general meaning given to the concept, but also about the on-going law projects, where the term everyman was used. The shift of focus from the general categories and their shares per decade (Table 4) to the yearly relative occurrences (Table 5) enabled us to move between these two levels. In contrast to topic modelling methods, which require well founded model parameters and explanation of the word cluster interpretation, cf. e.g. [14], FiST provided efficiently a description of the corpus through the semantic categories. The usefulness of FiST for the analysis was in the description of the semantic content, which appeared particularly applicable when comparing corpora and identifying differences and similarities. However, close reading of the relevant categories identified by FiST was a necessary step for validating the results and controlling the problems related to classification, and helped to engage with actual political discussion, which was not captured by the general semantic categories of FiST. The next steps in developing the FiST tool regard the management of the unmatched words category (Z99), which would improve the comparative analysis of historical texts. This could be achieved by improving the OCR quality of the parliamentary proceedings e.g. with automatic post correction of the texts and improving lexical coverage of FiST.

Acknowledgements

The work of the first author has been supported by the European Union's Horizon 2020 research and innovation program under grant 770299 (NewsEye). The analyses of the parliamentary data were run in the Taito cluster of the CSC - It Center For Science Ltd.

References

1. Kettunen, K.: FiST – Towards a Free Semantic Tagger of Modern Standard Finnish. IWCLUL2019, http://aclweb.org/anthology/W19-0306 (2019).
2. Koskenniemi, K. et al.: The Finnish Language in the Digital Age. META NET White paper series. http://www.meta-net.eu/whitepapers/e-book/finnish.pdf/view?searchterm=Finnish (2012).
3. Sahlgren, M.: The distributional hypothesis. Rivista di Linguistica 20.1, 33–53 (2008).
4. Erk, K.: Vector Space Models of Word Meaning and Phrase Meaning: a Survey. Language and Linguistics Compass 6/10: 635–653 (2012).
5. Tuunanen, P., Tarasti, M.: Everyman's rights and the code of conduct on private land. Finnish Ministry of the Environment, http://hdl.handle.net/10138/159060 (2015).
6. Sténs, A., Sandström, C.: Allemansrätten in Sweden: A Resistant Custom. Landscapes 15(2), 106–18 (2014).

7. Valguarnera, F. Accesso alla natura tra ideologia e diritto. Comparazione e cultura giuridica; 21. Giappichelli, Turin (2010).
8. La Mela, M.: Property Rights in Conflict: Wild Berry-Picking and the Nordic Tradition of Allemansrätt. Scandinavian Economic History Review, 62(3), 266–289, DOI: http://dx.doi.org/10.1080/03585522.2013.876928 (2014).
9. Matilainen, A.: Feelings of psychological ownership towards private forests. Doctoral dissertation, University of Helsinki, Faculty of Agriculture and Forestry, retrieved from http://hdl.handle.net/10138/300433 (2019).
10. La Mela, M.: The Politics of Property in a European periphery: The ownership of books, berries, and patents in the Grand Duchy of Finland 1850–1910 (Doctoral dissertation, European University Institute, Florence), DOI: http://dx.doi.org/10.2870/604750 (2016).
11. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation, MT Summit 2005, http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf (2005).
12. Fišer, D., Eskevich, M., de Jong, F. eds.: ParlaCLARIN: Creating and Using Parliamentary Corpora. ParlaCLARIN 2018 Workshop Proceedings. http://lrec-conf.org/workshops/lrec2018/W2/pdf/book_of_proceedings.pdf (2018).
13. Rouces, J., Borin, L., Tahmasebi, N.: Political Stance Analysis Using Swedish Parliamentary Data. In Digital Humanities in the Nordic Countries, Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, pp. 376–386 (2019).
14. Gentzkow, M., Shapiro, J.M. Taddy, M. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica* 87 (4), 1307-1340, https://doi.org/10.3982/ECTA16566 (2019).
15. Nanni, F., Osman, M., Cheng, Yi-Ru, Ponzetto, S.P., Dietz, L.: UKParl: A Data Set for Topic Detection with Semantically Annotated Text. In Fišer, D., Eskevich, M., de Jong, F., (eds.), pp. 29–32 (2018).
16. Blei, D. M.: Probabilistic topic models. Communications of the ACM 55:4, 77–84 (2012).
17. Nelimarkka, M. Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: kriittisiä havaintoja. Politiikka, 61(1), 6–33 (2019).
18. Tuulentie, S., Rantala, O.: Will 'Free Entry into the Forest Remain?' In D. K. Müller, L. Lundmark, H. Raynald (eds.), New Issues in Polar Tourism, pp. 177–188. Springer Netherlands (2013).
19. Sandell, K., Svenning, M. Allemansrätten och dess framtid : utredning om allemansrätten. Stockholm: Naturvårdsverket (2011).
20. Silfverberg, M., Ruokolainen, T., Lindén, K. et al.: FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. Lang Resources & Evaluation 50, 863–878, https://doi.org/10.1007/s10579-015-9326-3 (2016).
21. Löfberg, L.: Creating large semantic lexical resources for the Finnish language. Lancaster University, 422 pages (2017).
22. Multilingual USAS, https://github.com/UCREL/Multilingual-USAS
23. McArthur, T.: Longman Lexicon of Contemporary English. Longman, London (1981).
24. Navigli, R.: Word Sense Disambiguation: A Survey. ACM Computing Surveys 41,10–69 (2009).
25. Wilson, A. and Thomas, J.: Semantic annotation. In Garside, R., Leech, G. and McEnery, T. (eds.), Corpus annotation: Linguistic information from computer text corpora pp. 53–65. Longman, New York (1997).
26. Dullieva, K.: Semantic Fields: Formal Modelling and Interlanguage Comparison. Journal of Quantitative Linguistics, 24(1), 1–15, DOI: 10.1080/09296174.2016.1239400 (2017).
27. Geeraerts, D.: Theories of Lexical Semantics. Oxford University Press, Oxford (2010).

28. Piao, S., Archer, D., Mudraya, O. Rayson, P. Garside,R., McEnery, T. ,Wilson, A.: A Large Semantic Lexicon for Corpus Annotation. In Proceedings from the Corpus Linguistics Conference Series on-line e-journal, Vol. 1, no. 1, ISSN 1747-9398 (2005).
29. Suomen valtiopäiväasiakirjat. Eduskunnan kirjasto, Helsinki. 2012.
30. Kettunen, K.: Kirjoitetun nykysuomen automaattisesta semanttisesta merkitsemisestä. In Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen and Katja Västi (eds.), Proceedings Of The Research Data And Humanities (Rdhum) 2019 Conference: Data, Methods And Tools. Studia humaniora ouluensia, pp. 215–228 (2019).
31. La Mela, M.: Tracing the Emergence of Nordic Allemansrätten through Digitized Parliamentary Sources. In Paju, P., Oiva, M, and Fridlund, M. (eds.) Digital Histories. Emergent Approaches within the New Digital History. Accepted (2019).
32. Kettunen, K., Pääkkönen, T.: Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) http://www.lrec-conf.org/proceedings/lrec2016/summaries/17.html (2016).
33. Pirinen. T.: Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. SKY Journal of Linguistics, vol. 28, pp. 381–393, http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Pirinen.pdf (2015).
34. Klebanov, B.B, Diermeier, D., Beigman, E.: Automatic Annotation of Semantic Fields for Political Science Research. Journal of Information Technology & Politics, 5(1), 95–120 (2008).
35. Rayson, P.: From key words to key semantic domains. International Journal of Corpus Linguistics 13(4), 519–549 (2008).
36. Gabrielatos, C.: Keyness Analysis: nature, metrics and techniques. In Taylor, C., Marchi, A. (eds.) Corpus Approaches to Discourse: A critical review. Oxford: Routledge, pp. 225–258 (2018).
37. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999).
38. Franz, A., Brants, T.: All Our N-gram are Belong to You. https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html (2006).
39. Lison, P., Kutuzov, A.: Redefining Context Windows for Word Embedding Models: an Experimenta Study. In Proceedings of the 21st Nordic Conference of Computational Linguistics, pp. 284–288 (2017).
40. Niemikorpi, A.: Suomen kielen sanaston dynamiikkaa. Acta Wasaensia 26 (1991).
41. Hakulinen, A., Karlsson, F., Vilkuna, M.: Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus. Publications of the Department of General Linguistics, University of Helsinki, No. 6 (1980).
42. de Bolla, P., Jones, E., Nulty, P., Recchia, G, Regan, J.: Distributional Concept Analysis. A Computational model for History of Concepts. Contributions to the History of Concepts 14(1), 66–92 (2019).
43. Brofeldt, P.: Jokamiehen kalastusoikeus ja muut poikkeukselliset kalastusmääräykset. Otava, Helsinki (1943).