







Using parsed and annotated corpora to analyze parliamentarians' talk in Finland

Mykola Andrushchenko¹ | Kirsi Sandberg¹ | Risto Turunen¹ |
 Jani Marjanen²  | Mari Hatavara¹ | Jussi Kurunmäki¹  |
 Timo Nummenmaa¹  | Matti Hyvärinen¹  | Kari Teräs¹ |
 Jaakko Peltonen¹  | Jyrki Nummenmaa¹ 

¹Tampere University, Tampere, Finland

²University of Helsinki, Helsinki, Finland

Correspondence

Jyrki Nummenmaa, Tampere University,
 Kalevantie 4, 33100 Tampere, Finland.
 Email: jyrki.nummenmaa@tuni.fi

Funding information

Academy of Finland

Abstract

We present a search system for grammatically analyzed corpora of Finnish parliamentary records and interviews with former parliamentarians, annotated with metadata of talk structure and involved parliamentarians, and discuss their use through carefully chosen digital humanities case studies. We first introduce the construction, contents, and principles of use of the corpora. Then we discuss the application of the search system and the corpora to study how politicians talk about power, how ideological terms are used in political speech, and how to identify narratives in the data. All case studies stem from questions in the humanities and the social sciences, but rely on the grammatically parsed corpora in both identifying and quantifying passages of interest. Finally, the paper discusses the role of natural language processing methods for questions in the (digital) humanities. It makes the claim that a digital humanities inquiry of parliamentary speech and interviews with politicians cannot only rely on computational humanities modeling, but needs to accommodate a range of perspectives starting with simple searches, quantitative exploration, and ending with modeling. Furthermore, the digital humanities need a more thorough discussion about how the utilization of tools from information science and technologies alter the research questions posed in the humanities.

1 | INTRODUCTION

The massive production of born-digital texts as well as digitalization of historical records has opened up new avenues of study, making it possible to analyze machine-readable collections. Among the types of

historical sources available, newspapers and parliamentary records have most commonly been transformed into digital form. They stand out as they form a long series of fairly uniform textual data. While the use of newspapers is often restricted due to copyright issues, parliamentary records are public and free from

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Journal of the Association for Information Science and Technology* published by Wiley Periodicals LLC on behalf of Association for Information Science and Technology.

any copyright restrictions that could hamper their use for research purposes. Hence, we can find digital collections of them from most European countries. For parliamentary records, the European CLARIN Infrastructure that collects, maintains, and develops common language resources links to 36 different resources that contain European linguistically annotated parliamentary records.

Due to their institutional and procedural similarities and a long history of public records, parliamentary documents provide an easily accessible collection of sources with great potential for national, transnational, and comparative research (Ihalainen & Palonen, 2009). Parliament is a crucial research interest for many disciplines ranging from politics, history, and sociology to the study of rhetoric. Parliaments have for a long time been the main object of struggles for democratization (Bonin, 2020; Ihalainen, Ilie, & Palonen, 2016; Kurunmäki, Nevers, & te Velde, 2018) and have also received attention from a digital humanities perspective (see, e.g., Curran et al., 2018; Grimmer, 2013). Today, parliament is the main forum for the nationwide political debate and has a formalized link to political decision making in democratically governed countries (e.g., Ilie, 2010). Parliamentary debates both reflect upon and feed political debates in the press and social media, and the society at large.

The Finnish parliamentary records and the accompanying interviews have so far been used only in a few studies that deploy methods from corpus linguistics, language technology, or computer science (Kettunen & La Mela, 2020; Loukasmäki & Makkonen, 2019; Nelimarkka, 2019). In the past few years political scientists have also started to discuss the broader implications in research culture, collaboration, and research questions when moving toward computational methods in the analysis of parliamentary material (Ahonen, 2015; Ahonen, 2018; Ahonen & Wiberg, 2018). In particular, one of the problems with analyzing parliamentary speech quantitatively lies in the tension between the inherent activity of speaking for and against (*pro et contra*) in parliament, and the flattening of conflict that often follows from quantifying linguistic features.

The two datasets have been grammatically parsed for the purpose of analyzing linguistic features to answer humanities questions. The Finnish language has a rich system of inflections. Finnish words may appear in many distinct inflectional forms, including fairly rare ones. This makes Finnish poorly suitable for direct machine learning or bag-of-words models. Additionally, Finnish is a relatively small language in terms of resources compared to languages like English and Spanish, which makes straightforward machine-learning even more challenging. A common remedy for this is to start by parsing the texts, for

example, into Universal Dependencies (UDs) (Universal Dependencies Website, <http://universaldependencies.org>).

Text corpora allowing access to grammatical elements of the text exist, with contents such as the Finnish parliamentary records 2008–2018, like the Korp corpus offered by Finnish Language Bank Kielipankki (Plenary Sessions of the Parliament of Finland, 2020). That corpus can be searched using an intuitive graphical search interface or using the relatively general CQP (corpus query processor) query language. Building on this development, our work is based on implementing our own search system and corpora (VoDe Corpora, 2020), allowing us to make queries that explore larger passages of text and add flexibly new content to the corpora.

In addition to our system, we introduce three case studies as three different humanities questions all benefitting from the grammatical parsing in the two corpora. A case on discourses of power among politicians is demanding to interpret because it requires a good understanding for when interviewees talk about power and especially when they do not. Lemmatization and parsing is central to finding relevant passages and for understanding when discourses of power appear. Our case on isms as pivotal terms for organizing political thinking highlights the need for solid quantification of different, but related, terms across the political spectrum. It targets co-occurrences and chronological shifts in political language, but ultimately cannot interpret quantitative results without toggling back and forth between textual examples and figures. The third case moves the focus to a central concept of humanistic interpretation: narrative. We use grammatical tenses to identify narrative structures in the text tackling a computationally challenging task that has great potential for transferability. The cases illustrate the need for combining easy access for reading and exploring the data for humanities scholars with the possibility of tailoring particular analyses to answer specific research questions.

While parliamentary records provide an exceptionally good resource for studying language and using methods of Natural Language Processing (NLP), the biggest interest toward parliamentary speech has been in the disciplines that focus explicitly on politics. The use of grammatical parsing, other linguistic annotation and methods developed for NLP to answer questions that are ultimately about understanding politics and political discourse is not always straightforward. Hence, this article turns to the process of organizing and cleaning the data, using language resources to analyze the data and asking humanities and social science questions to the data. In doing so, it harmonizes with earlier discussions about collaboration between computer science and humanities researchers (see Biemann, Crane, Fellbaum, & Mehler, 2014; Crum, Angello, Liu, & Campion, 2019; Mäkelä

et al., 2019) and intervenes in the current debate about the nature of digital humanities and its relationship to computational modeling. Currently, the debate about the future of digital humanities has revolved around its extremely broad scope and variety of methodological choices. Some have argued for a separation of “contemporary humanities,” that is analyses assisted by new methods and “computational humanities” that consist of studies that choose humanities phenomena and try to model them computationally (see especially, Piotrowski, 2019, 2020). Our three cases are all ambitious attempts in making a humanities impact, but their level of ambition with regard to computation varies a lot. Still, taken together they indicate that a neat separation between modeling in computational humanities and exploration in contemporary humanities does not fit well. Even the computationally heaviest example, the one about modeling narrative, can only be made intelligible by combining the two perspectives.

2 | THE CORPORA AND THE SEARCH SYSTEM

This paper is based on analysis of parliamentary records and oral history interviews with previous MPs. A project of the Finnish parliament has interviewed experienced veteran MPs since 1988. Our database contains all available and transcribed Finnish-language interviews. The interviews are transcribed by using the same norms as transcribing the parliamentary records, which means that the interviews are recorded in standard Finnish. The second dataset has been drawn from transcripts of plenary sessions of the Finnish parliament from February 1980 to November 2018. Official records of the Finnish parliament include some editorial changes with regard to regionalisms, spoken variants, some syntactic variants, self-corrections, and selected particles (Voutilainen, 2017). The records were stored as structured XML files or textual PDF files. Each transcript file contained the speeches made by members of parliament (MPs), and related metadata such as the date and time, the issue under discussion, the speaker's party affiliation, and whether the speech was a reply. The metadata was stored and associated with individual sentences.

The interviews with veteran members of the Finnish parliament are long semistructured conversations based on a fixed selection of subjects, including the parliamentarian's background and biography; election campaigns; everyday work in the parliament; relationships with significant figures such as the prime minister; and personal views, for example, on the purpose and role of the parliament. Former ministers or MEPs were asked additional questions about

these aspects of their careers. The scope of the Finnish-language data made available for research use is 378 interviews, roughly 883,000 sentences and 11,890,000 word tokens (except punctuation). Corresponding interview collections exist at least in the UK and the United States, yet the collection is both rather unique and sparsely used because of its prominent size.

Parliamentary records include various announcements made on different occasions during plenary sessions, always employing the same phrasing for consistency. An example of such an announcement is a vote held about a particular bill and the announcement of its results. These contents were retained in the documents for consistency, but excluded from further analysis. The dataset thus contained only the MPs' transcribed utterances, with a total of roughly 5,200 sessions, 5,285,000 sentences and slightly over 80 million word tokens.

2.1 | Data processing for the search system

All textual information was parsed and annotated by the Finnish Dependency Parser, TurkuNLP (Haverinen et al., 2014), producing Universal Dependencies (UDs). Morphological specifications of words in UD schemata contain (a) lemmas representing the semantic content of the words, (b) part-of-speech tagging representing the abstract lexical categories, and (c) features representing lexical and grammatical properties that are associated with the particular words. The sentences are represented as trees with words as nodes and their grammatical relationships as edges connecting the nodes. Such a representation obviously carries more accurate information of the text content than treating the text as just a sequence of a set of words. Syntactic information for each word featured its function from the list of UD relations, as well as a single word that the current word depended on.

Morphological identification is particularly beneficial when dealing with languages that have a plethora of inflectional forms, carrying grammatical information, as in the Finnish language. It enables the capture of different inflected forms under the same root form, regardless of how diversely they appeared in the texts. Conversely, it allows extensive and precise searches for individual forms without retrieving irrelevant results.

Texts from various datasets were processed by separate pipelines, depending on their original format and quality, and converted into the same structure. Sentences were morphologically parsed, extended with UD interpretations, and stored with dataset-specific metadata, discussed above. Each sentence was stored in a MongoDB collection as an individual document. This database

system was chosen thanks to its flexible schema and a wide variety of query operations; the datasets were also small enough to achieve reasonable performance.

2.2 | The search tool

A custom search tool has been developed as a web application to search the dataset. Supporting regular expression syntax, it enables the user to retrieve individual words and their combinations according to dictionary forms, with additional filters on arbitrary morphological properties. Specifying the first few letters of a word is a simple way to find all words with the same root, while the last few letters can capture all words derived using the same ending. This also interacts well with the common Finnish word formation mechanism, compounding. For example, searching for “work” can also retrieve such words as “work shift,” “workplace,” and “work contract,” or conversely “part-time work,” “overtime work,” and “shift work.” These are all single words in Finnish, parsed in their entirety.

The word component of the search allows the user to locate an individual word, a fragment of it or an entire class of words, whether in any inflected form, only in the specified form or only with a certain morphological property (e.g., past-tense verbs or singular nouns). Separate word-specific queries can be joined by AND or OR operators, resulting in more complex queries. Metadata filters can also be applied to the search results, further restricting them in terms of timeframe, speaker or

political party. This results in a powerful search mechanism, although the cost of such flexibility is a certain steepness of the learning curve.

The basic operation of the search tool is demonstrated in Figure 1, showing the lemma string in the upper-left corner: all nouns starting with “work,” in any plural form. Metadata filters are found directly below that textbox and several matching sentences are presented below. A translation of the last entry is provided in the dashed box at the bottom of the figure; this does not appear in the actual system. Checkboxes on the right control the visibility of individual columns, which is useful to remove unnecessary details from view.

The final part of the query (*Number = Plur*) applies to the morphological annotation produced by the dependency parser. Figure 2 shows such an annotation for sentence 1776/3464 of Figure 1. Here, *form* is the actual form encountered in the sentence, *lemma* is the inferred basic form, while *head* and *deprel* represent UD information. The 17th and the 19th word in Figure 2 match the beginning of the query, but only the 17th is a plural form and thus fully satisfies the request. Any property listed in the *feat* column may be incorporated into the query as a filter.

Context can be requested by the user, yielding a number of extra sentences immediately before and after each search result. The tool also implements a “window search” mode for more sophisticated context retrieval. In this mode, the search begins from the first keyword of the query and checks the sentences

The screenshot shows the search tool interface. At the top, there are input fields for 'Lemma' (with a dropdown menu showing 'työ|Type=NOUN|Number=Plur'), 'From year' (2000), and 'to year' (2010). Below these are 'General filters' including 'Party' (No restriction) and 'Sentence source' (No restriction). 'Dataset-specific filters' include 'Language' (Any) and 'Speaker'. 'Context settings' include 'Return' (0) and 'Maximum distance between sentences' (0). On the right, 'Column visibility controls' are shown as a list of checkboxes for columns: party, sp_num, p_num, session, item, agenda, day, date, language, paragraph, and type. Below the filters is a search bar and a search button. The search results are displayed in a table with columns: sentence, text, name, surname, status, party, sp_num, session, agenda, day, date, language, paragraph, and type. The table shows four results. The last result (2312/3464) is highlighted with a dashed border, and its translation is shown in a dashed box below it.

sentence	text	name	surname	status	party	sp_num	session	agenda	day	date	language	paragraph	type
1480/3464	Toinen yhtä suuri on työsopimuslain uudistus, joka nyt on venynyt jo minusta turhan pitkään ilman ratkaisua, koska on tehty selvityksiä, että tuolta rajojen takaa jopa 400 000 ihmistä olisi sen jälkeen, kun lahden takana ollaan EU:n jäsenvaltioita, valmiita tulemaan Suomeen töihin.	Matti	Huutola	edustaja	vas	553	99	Hallituksen esitys valtion talousarvioksi vuodelle 2001	ti	12.9.2000	suomi	380	vastaus
1776/3464	Sekä tutkimukset että terve järki sanovat, että paras tulos saadaan aikaan alentamalla sekä tuloverotusta että työntantajamaksuja pienipalkkaisen työn osalta.	Maria	Aula	edustaja	kesk	380	99	Hallituksen esitys valtion talousarvioksi vuodelle 2001	ti	12.9.2000	suomi	452	varsinainen
1911/3464	Tekemättömien töiden jatkuva kuormitus ja kiire vievät voimat ja mielekkyyden alan ammattilaisilta.	Mikko	Immonen	edustaja	vas	456	99	Hallituksen esitys valtion talousarvioksi vuodelle 2001	ti	12.9.2000	suomi	480	varsinainen
2312/3464	Hallitus ei myöskään ole juurikaan keventämässä työntajien sivukulurastetta, vaikka se olisi hyvä ja kestävä keino työn kannustavuuden lisäämiseksi.	Sakari	Smeds	edustaja	skl	500	99	Hallituksen esitys valtion talousarvioksi vuodelle 2001	ti	12.9.2000	suomi	571	varsinainen

2312/3464	The government is not at all reducing the strain of employers' incidental expenses either, although that would be a good and sustainable way of improving work attractiveness.	Sakari	Smeds	MP	skl	500	99	Government proposal of the state budget for the year 2001	Tue	12.9.2000	Finnish	571	proper
-----------	--	--------	-------	----	-----	-----	----	---	-----	-----------	---------	-----	--------

FIGURE 1 A simple search [Color figure can be viewed at wileyonlinelibrary.com]

id	form	lemma	cpos	feat	head	deprel
1	Sekä	sekä	CONJ		2	cc:preconj
2	tutkimukset	tutkimus	NOUN	Case=Nom Number=Plur	6	nsubj
3	että	että	CONJ		2	cc
4	terve	terve	ADJ	Case=Nom Number=Sing Degree=Pos	5	amod
5	järki	järki	NOUN	Case=Nom Number=Sing	2	conj
6	sanovat	sanoa	VERB	VerbForm=Fin Number=Plur Person=3 Tense=Pres Voice=Act Mood=Ind	0	root
7	,	,	PUNCT		11	punct
8	että	että	SCONJ		11	mark
9	paras	hyvä	ADJ	Case=Nom Number=Sing Degree=Sup	10	amod
10	tulos	tulos	NOUN	Case=Nom Number=Sing	11	dobj
11	saadaan	saada	VERB	Voice=Pass Tense=Pres Mood=Ind VerbForm=Fin	6	ccomp
12	aikaan	aika	NOUN	Case=Ill Number=Sing	13	nmod
13	alentamalla	alentaa	VERB	Case=Ade Voice=Act InfForm=3 Number=Sing VerbForm=Inf	11	advcl
14	sekä	sekä	CONJ		15	cc:preconj
15	tuloverotusta	tulo#verotus	NOUN	Case=Par Number=Sing	13	dobj
16	että	että	CONJ		15	cc
17	työnantajamaksuja	työnantaja#maksu	NOUN	Case=Par Number=Plur	15	conj
18	pienipalkkaisen	pieni#palkkainen	ADJ	Case=Gen Number=Sing Degree=Pos	19	amod
19	työn	työ	NOUN	Case=Gen Number=Sing	20	nmod:poss
20	osalta	osa	NOUN	Case=Abl Number=Sing	13	nmod
21	.	.	PUNCT		6	punct

FIGURE 2 Word-by-word morphological annotation [Color figure can be viewed at wileyonlinelibrary.com]

1692-1693/2054	Euroopan unionia tulee kehittää kansallisvaltioiden ja hallitusten välisenä liittona, ei suinkaan liittovaltioksi. EU-parlamentin valtaa ei tule lisätä kansallisten parlamenttien kustannuksella, vaan Suomen tulee yhdessä muiden Pohjoismaiden kanssa puolustaa kansanvaltaista yhteiskuntajärjestelmää ja kansallisen itsemääräämisoikeuden vahvistamista.	Leea	Hiltunen	edustaja	skl	401	4	ke	9.2.2000	suomi	varsinainen
957-959/1044	Kyllä toisaalta massatyöttömyys syntyi Ahon hallituksen aikana. Silloinhan lähti liike siihen suuntaan, että eriarvoisuus syntyi. Näyttää siltä, että tämä kehitys vielä edelleen jatkuu, mutta kyllä se pantiin liikkeelle silloin, kun Ahon hallitus oli vallassa .	Unto	Valpas	edustaja	vas	604	10	ke	16.2.2000	suomi	varsinainen
216-219/557	Hallituksen esitys olisi ollut eräitä ratkaisevita perusteillaan parempi kuin lakivaliokunnan esitys. Toki lakivaliokunnan mietinnössä on parannuksiaikin hallituksen esitykseen verrattuna, mutta mietintö antaa aivan liian suuren vallan korkeimmalle oikeudelle ja korkeimmalle hallinto-oikeudelle. Nyt ei korkeimmille oikeuksille edes riittänyt se, että ne saavat itse valita omat edustajansa tuomarinvallintalautakuntaan, vaan lakivaliokunta muutti hallituksen esitystä niin, että mainitut tahot saavat nimetä muidenkin tahojen edustajat poisluehtuna kuitenkin syyttäjät, asianajajat ja tieteen edustajat. Hallituksen esityksessä korkein oikeus ja korkein hallinto-oikeus olisivat saaneet antaa vain lausunnon lautakunnan asettamisesta.	Kari	Myllyniemi	edustaja	kesk	530	15	pe	18.2.2000	suomi	varsinainen
348-350/557	Niin löytyi tämä kompromissi, että sidotaan enemmän kuin hallitus esitti niitä jäseniä, jotka lautakuntaan on nimettävä. Kun lisättiin lautakunnan jäsenmäärä 12:een, tultiin siihen, että valtioneuvostolla ei ole kuin kolmen neljän kohdalla jonkin verran harkintavaraa siinä, ketkä siihen nimitetään jäseniksi. Mielestäni tämä kompromissi on sellainen, että se ei anna periksi hallitukselle eikä niille tahoille, jotka vaativat koko lautakuntajärjestelmän unohtamista, vaan antaa oikeuslaitokselle lisää valtaa ja sitoo hallituksen käsiä enemmän kuin hallituksen esitys.	Toimi	Kankaanniemi	edustaja	skl	175	15	pe	18.2.2000	suomi	varsinainen
471-473/557	Useat ovat puoltaneet hallituksen esitystä, jossa valtioneuvostolla on kyseisen lautakunnan asettamisvalta. Useat kuultavana olleet asiantuntijat ovat puolestaan korostaneet, että lautakunnan asettamisvalta kuuluu tuomioistuimilaitokselle, sillä tämä käytäntö ilmentäisi tuomioistuimen riippumattomuutta hallitusvallasta. Katsonkin, että on tärkeää, ettei valtioneuvostolle anneta eräänlaista kaksinkertaista valtaa tuomareiden nimittämiskäytännössä; sekä tuomarinvallintalautakunnan asettamisessa että ratkaisuehdotuksen tekemisessä tasavallan presidentille.	Leena-Kaisa	Harkimo	edustaja	kok	562	15	pe	18.2.2000	suomi	varsinainen

FIGURE 3 Window search

adjacent to every match, up to a specified distance from it. As long as subsequent keywords are found nearby, the “window” expands backward and forward from them; if every keyword is present, the entire captured block of text is returned to the user. Figure 3 shows this mode in action, with highlighted keywords (“government” and “power”) located at most two sentences away from each other. The keywords may appear in different inflected forms.

The combination of metadata, word forms and morphological properties can be used to retrieve initial findings for a number of research questions. The task of the

user is to formulate a reasonable query, not overlooking relevant words or their groups, and, when necessary, adjust the query to reduce the result set without losing relevant results.

A basic log of search activity was maintained, partly to assist in troubleshooting individual queries when necessary. The log provided information about which user activated the tool, what search terms were included in the query and what filters or search parameters were in use. These details were also provided to maintain a personalized search history: a long list of each user’s most recent queries was available to them, and selecting an

entry from that list restored the search form to the same values as used in the “historical” query. Thus, it was possible to return to a successful query and modify it further.

2.3 | Users and usage of the tool

Users of the tool were researchers from different fields: ten experts in history of concepts, Finnish history, political science, narratology, sociology, linguistics, digital humanities, statistics, and data mining. A specialist from the Library of Parliament also participated in the project as a coordinator of the interview dataset. These researchers started their work at different stages of the project. They utilized the tool for their own research, at different times and in various ways, depending on their ongoing projects and needs.

Within about two and a half years since the first deployment of the tool, it has been executed just under 2,800 times. This figure does not include queries made by the developer: he mostly attempted them for testing purposes, often repeating the exact same search many times until obtaining the desired results. Most “regular” users performed between 150 and 200 searches.

Analysis of the log file, especially the lemma components of the searches, presents certain suggestions about the thought patterns of the users and their ways of making searches with the tool. A brief instruction was drafted shortly after the first version of the tool, showcasing the most common search elements and their combinations. Furthermore, the small size of the project team allowed the developer to participate in meetings with the researchers regularly. He could formulate exact search queries based on individual needs, modifying them further as requested, and it was hoped that the same patterns could be generalized to further searches.

A fairly common search strategy is using a series of queries, produced within seconds or minutes, with largely the same parameters and only minor changes in the lemma string. The first attempts are often incorrect or needlessly restrictive, suggesting that the user made adjustments until sufficiently many results were returned. However, it also appears that these adjustments were sometimes made randomly, in order to discover something that happened to work, rather than as a conscious effort to form a proper query.

This effect can be attributed to the differences between our tool and other popular search mechanisms. The tool relies heavily on regular expression syntax without stepping away from it for reasons of convention. Searching for a word fragment already returns longer words with the same substring: extra characters are not needed to require it specifically, but rather to prevent

these longer words from appearing. Common AND and OR operators are not used and the operands are instead grouped using quotation marks, which accordingly do not have an “exact form” connotation. And while searching for a basic form generally returns all the possible inflections, searching for an exact inflection has no effect—again, unless this exactness is explicitly required.

One fully justified scenario with repeating search terms could be observed when the scope of requested context changed between the queries. For example, a user might have first attempted to retrieve just one extra sentence, then repeated the same search with three. Presumably, the context fetched by the first search was sometimes attractive but not sufficient, so that more text was needed around certain hits. The impact of the “radius” parameter is even greater in the window search mode, where increasing it may yield completely new results. This is because, for instance, keywords located within three sentences of each other will not be retrieved by a search with a maximum distance of two.

Another frequent occurrence was a sequence of searches within a single session and with the same parameters, but substantially different keywords (nonetheless belonging to a certain theme or simply related, for example, various words ending with “ism”). The keywords could be arranged in one query as OR operands, yet this was not necessarily desired. This pattern likely illustrates the search tool’s explorative potential, when a look through the material already suggests new terms for further searches.

Finally, the general trend over time is that searches have gradually become somewhat more advanced. Some functions remain clearly unpopular, such as keyword negation that excludes results featuring a particular word: apparently, used keywords are often rare enough to require no further elimination. However, filters on morphological properties and grouped terms have been used with more consistency, and this is not only a consequence of agreeing queries with the developer but of independent search-making and learning. Likewise, when a chain of improper searches is found, more often it tends to evolve toward a reasonable query rather than trailing off fruitlessly.

Certain lessons can definitely be learned from the queries accumulated so far, such as the role of conventions and the tradeoff between flexibility and simplicity.

3 | CASE STUDIES

This section presents three different scenarios where a research problem was approached by specialists from

various fields, aided by the developed search tool: an investigation into the understanding of power, a study of “isms” in a temporal context, and an attempt to identify narratives within the data. The approaches chosen in each case differ notably from each other and also result in fairly diverse utilization of the search tool, exposing its strengths and weaknesses. References made in the description of each scenario only apply to that particular case.

3.1 | The self-understanding of power

Political scientists, sociologists and philosophers have pursued the most accurate definitions of “power.” This study, resorting to the history of concepts approach, asks instead how the actual political actors, in this case former members of the Finnish parliament, use the concept of power in their oral history interviews. In doing so, we pose the following questions: (a) How often is the term “power” used, absolutely and in comparison with other key concepts? Are there temporal trends in the numbers between 1988 and 2018? (b) Is power understood as a constant power-of-command, a power of decision, belonging to separate individuals, or rather as a chance requiring cooperation, skills and proper timing? (c) Is power conceived as a structural (e.g., belonging to a class), collective (e.g., belonging to a party) or individual phenomenon? (d) How is power evaluated (e.g., in negative, positive, or critical terms)? (e) How stable or context-dependent concepts do individual veteran MPs have?

The focus of this study is on qualitative political text analysis. The results of earlier research, based on smaller, hand-picked materials, suffer from low generalizability (Hyvärinen, 2003). The available Finnish-language interviews comprised almost 12 million words—far too much for traditional qualitative analysis, for reading, selecting and marking individual hits, let alone counting and comparing the frequencies. The interviews in our corpus were searched with the lemma of “valta” (power). Each located sentence was accompanied with metadata, including the name and political party of the interviewee, the date of the interview, and the name of the interviewer. This often enabled further reasoning based on the political biography of the interviewee and the context in which power was mentioned.

The mechanism of word compounding added a complication to the computational analysis. A common Finnish way in coining new terms is to build compound nouns. Being a short and old word, “valta” (power) is used to derive many other concepts. These include both unrelated ones, such as violence (“*väkivalta*”) and realm (“*valtakunta*”), and fairly relevant ones, for example,

authority (“*arvovalta*”) and influence (“*vaikutusvalta*”). The irrelevant compounds were explicitly specified in the queries. The corpus of relevant search results required cleaning and reflection concerning the conceptual limits. In our case, we omitted violence but included authority and influence. Our argument is that, even though power and authority are clearly distinct concepts in English, the Finnish form of the words (“*arvovalta*,” “*vaikutusvalta*”) invites them to a particular “valta-family” of terms.

Finnish language is easier to search than English in the sense that some common uses of English “power” (e.g., electric power) are not expressed with “valta” in Finnish. Literally, “nuclear power” is “*ydinvoima*,” “nuclear force,” in Finnish. Our study analyzes the ways politicians use the language of “valta” (power). When the former MPs discuss power issues without using this concept is an entirely different question, which could only be answered by qualitative analysis.

The whole corpus contained 3,506 references to power (“valta”), which makes it the 98th on the list of most popular nouns. Among the other key concepts of politics, power had a moderate position. The most common concepts of politics were (political) party (25376), government (20251), politics (10609), and state (3703). The interviewed veteran politicians preferred concrete concepts over more abstract terms, such as citizen (1506) or democracy (1007). The quantitative analysis revealed an increase in the use of “valta” (power) over time. However, closer analysis showed that the language of power was typically used only when prompted by the interviewer. Interviews were conducted by 17 historians, and a few of them had a great interest in issues of power. In the early 1990s, one of the interviewers noted the relevance of power and suggested a change in the thematic structure of the interviews. Even after the revision of the interview guide, the differences between interviewers endured. Rather than finding a solid, quantitative trend of the references of power, we discovered the crucial role of the interviewers and the interviews’ thematic structure (see Holstein & Gubrium, 1995). Tracing down the origin of the differences required contacting the actual interviewers.

After the basic counting, the amount of search hits was still high for careful qualitative analysis. At the same time, the qualitative questions resisted operationalizations that could have enabled further computerized analysis. The discrepancy, however, was not so much between digital and humanistic ways of study as between political thinking and linguistics, in the sense that the distinctions relevant for the qualitative analysis of concept usage did not have clear linguistic markers, which could have been captured by digital means. The only solution was seen in the laborious reading and marking of the 3,506 references, which was performed manually.

A close examination of the occurrences of power resulted in the following preliminary observations. (a) The concept of power was not used as often as suggested in previous literature (Haugaard & Clegg, 2009, p. 1). The amount of references was, to a relevant extent, dependent on the interview questions and the interviewers. (b) Part of the veterans privileged the understanding of power as personal power-of-command and power to decide. For them, the role of the MP did not provide any relevant power, in contrast to the positions in the Cabinet, city administration or trade unions. Those who focused on power in the parliamentary process often used the terms “vaikutusvalta” (influence) and “arvovalta” (authority). (c) Part of the most leftist interviewees systematically avoided discussing the power of an individual MP and refuted the power of the parliament as regards the “power of economy.” On the other side of the political spectrum, interviewees often resisted the “power of trade unions.” In particular, the veterans from previously leading government parties also had the perspective of the power of an individual MP, working in and between the parliamentary groups, committees, and party leadership. (d) The range of possible evaluations of power is wide. The discourse of corrupting power was used in particular by the opponents of President Urho Kekkonen (whose term “in power” continued from 1956 to 1982). However, most veterans accepted the necessity of power in pursuing desirable policies. (e) Rather than employing a single solid concept of power, most interviewees kept using several, even contradicting ways of using “power.” Often, one’s own actions were conceptualized in terms of advancing good policies and the opponents’ actions in terms of power. Therefore, power functioned more prominently as a concept of criticism and resistance than as a tool of neutral description.

Two significant conclusions can be made from the perspective of this case study. First, morphological distinctions have been only partially useful in obtaining valid search results: semantically different words related to the same concept may be constructed in the same manner, and the search tool was unable to distinguish relevant ones among them. Second, the “traditional” analysis of concepts was possible to this extent only after parsing the interview material in a way that enabled the efficient selection of the relevant statements.

3.2 | Isms in parliamentary speech

Ism-words such as “liberalism,” “republicanism,” and “communism” are forward-looking projections that have been crucial in the formation of modern political culture (Höpfl, 1983; Koselleck, 2011) and important in the

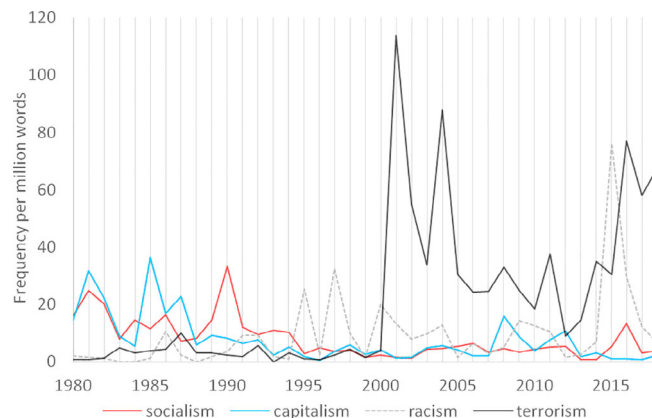


FIGURE 4 Relative frequency of four popular isms in the Finnish parliament, 1980–2018

construction of ideological, cultural, religious and scientific traditions (Kurunmäki & Marjanen, 2018b). Although many scholarly interventions claim that ideological grand narratives or doctrines can no longer guide and explain the political imagination and action (e.g., Lyotard, 1979), others debate whether new (or reactualised) isms, such as “populism,” “neoliberalism,” and “multiculturalism,” are proper ideologies with coherent doctrines and historical narratives of their own (e.g., Freeden, 2017).

In order to assess whether old isms still matter or whether new isms have taken their role as navigating concepts in political rhetoric, we have investigated the use of ism-words in Finnish parliament between 1980 and 2018 through the search system. We quantified the temporal variation of popular ism words in the Finnish parliament. Figure 4 shows that the classical isms of political positioning, that is, “socialism” and “capitalism,” had still been actively used in the 1980s, but their popularity decreased after the collapse of the Soviet Union in 1991. The September 11 attacks in the United States affected the use of isms in Finland by raising “terrorism” to the core of political debate. Interestingly, the effect persisted as terrorism remained frequently used in most years after 2001. Another new ism, “racism,” was first spoken of in 1986 and became a steady part of parliamentary debate in the 1990s onwards.

Using temporal metadata to plot relative frequencies of ism words over time indicated a long-term shift from “socialism” and “capitalism” to “terrorism” and “racism,” but time series analysis does not shed much light on the historical factors behind the shift. Thus, we decided to explore party-political metadata connected to ism words. At this point we did not discriminate against any isms, as the inclusion of even seemingly unideological isms such as journalism or realism is telling of the party profiles.

TABLE 1 Most popular isms of four parties in the Finnish parliament

Social democratic party 1980–1991		National coalition party 1980–1991	
Top 10 ism words	Relative frequency per million words	Top 10 ism words	Relative frequency per million words
Parliamentarism	71.7	Parliamentarism	22.7
Socialism	17.8	Socialism	20.2
Capitalism	13.4	Realism	7.0
Populism	12.6	Humanism	7.0
Realism	7.8	Optimism	5.4
Optimism	7.8	Populism	4.9
Racism	5.9	Communism	4.5
Facism	3.7	Terrorism	4.1
Terrorism	3.3	Capitalism	2.9
Market capitalism	3.3	Protectionism	2.9
Green league 1992–2018		Finns party 1995–2018	
Top 10 ism words	Relative frequency per million words	Top 10 ism words	Relative frequency per million words
Terrorism	43.9	Terrorism	59.1
Parliamentarism	25.4	Racism	19.2
Racism	21.8	Socialism	17.8
Realism	16.5	Parliamentarism	16.6
Populism	15.5	Realism	15.2
Journalism	5.3	Populism	14.2
Socialism	4.1	Capitalism	8.8
Optimism	3.9	Communism	7.8
Protectionism	3.4	Journalism	6.2
Everyday realism	3.2	Age racism	4.5

Table 1 shows the most frequent isms used by four different parties: first, in the upper part of the table, the two major parties in the Finnish parliament of 1980–1991, that is, moderate left-wing Social Democratic Party and moderate right-wing National Coalition Party, and second, in the lower part of the table, the two new parties that have gained popularity after the collapse of the Soviet Union: the liberal Green League and right-wing populist Finns Party.

The SDP still used “socialism” as the positive alternative to “capitalism” and “market capitalism” in the 1980s but later distanced itself from the concept that had been crucial for the ideological make-up of the party since its birth in 1899: the relative frequency of socialism dropped to a mere 4.3 instances per million words during 1992–2018. The National Coalition Party also constructed its political identity with “socialism” as a counter concept before, but switched from “socialism” to “communism” in the post-Soviet era: the relative frequency of socialism decreased to 4.9 instances per million words, whereas “communism” simultaneously rose to 13.2 per million

words, the highest number among all parties. The rising new parties, the Green League and the Finns were the most active to use both “terrorism” and “racism” in the parliamentary debates of 1992–2018. This does not mean that the new parties would explain the rise of new isms. Rather, all parties increased their references to “terrorism” and “racism” over time, but the new parties used them most frequently. It can be assumed that the old division of clearly competing ideological isms (e.g., socialism/communism versus capitalism) worked better for the traditional parties than the new era of unilaterally negative isms (e.g., terrorism, racism) which, in turn, favored the newcomers of Finnish politics.

It has been pointed out in previous studies of isms that these words often appear in clusters with other isms (Kurunmäki & Marjanen, 2018a). In order to see whether this is the case in Finnish parliamentary debates, we traced co-occurrences of isms using the “window search” feature of the tool with a distance of three. That is, the tool returned all instances (2019 in total) where two isms, same or different ones, appeared within

TABLE 2 Isms most commonly appearing close to other isms in the Finnish parliament, 1980–2018

Ism pair	Absolute frequency
Socialism and capitalism	63
Communism and Nazism	20
Socialism and communism	11
Optimism and pessimism	11
Parliamentarism and normal parliamentarism	9
Socialism and real socialism	7
Parliamentarism and populism	6
Nazism and fascism	5
Terrorism and racism	5
Fascism and communism	5
Optimism and realism	5

three sentences of each other. Such textual fragments were further expanded left and right until no more isms emerged for three consecutive sentences. Each ism of a given window formed a pair with every other one, and also with itself if it appeared multiple times.

As can be seen in Table 2, the far most common pair of isms is “socialism” and “capitalism,” followed by “communism” and “Nazism” and “socialism” and “communism.” While our findings may seem expected, they nonetheless help us develop more detailed research questions when combined with the knowledge gained from previous research and the possibility to expand the context provided by our search tool. It is therefore possible to make at least three research hypotheses.

First, based on the history-theoretical notion of antagonistic conceptual pairs as an important characteristic of political language (Koselleck, 2011), we argue that isms in Finnish political rhetoric form pairs of opposites such as “socialism” and “capitalism” or “parliamentarism” and “populism.” Second, given the notion that ideological isms have often been used as pejorative labels or regarded as harmful altogether (Kurunmäki & Marjanen, 2018b), it is clear that some of the conceptual pairs are not antagonistic but, rather, indicate some kind of similarity. Therefore, pairs such as “Nazism” and “fascism” but also “socialism” and “communism” or “communism” and “Nazism”—depending on the context and the speaker—are rhetorical equations in which negative sentiment is transferred between the isms. Third, pairs such as “parliamentarism” and “normal parliamentarism” as well as “socialism” and “real socialism” suggest that much of the discourse of isms has been about using them as navigating concepts and qualifying them through adjective use.

The study of isms used the search system in a twofold manner. First, it performed some exploratory data analysis in order to construct bottom-up hypotheses on the role of isms in the parliamentary debate. Temporal analysis helped scholars to generate a hypothesis on the diachronic change in the use of isms, party-political analysis showed that the change was not caused by an individual party but by several parties, perhaps indicating a shift in the Finnish political culture as a whole, and co-occurrence analysis showed that the classical ideologically laden isms (e.g., socialism and capitalism) clustered together more strongly and performed different rhetorical functions than the most popular isms of the 21st century (e.g., racism and terrorism). Second, it used the possibilities to produce simple statistics of normalized frequencies and co-occurrences to strengthen points arising from reading text passages in context. Producing statistics that were tailored to answer very specific humanities claims highlights the need for toggling back and forth between qualitative interpretation and quantitative representations of parliamentarians' talk.

3.3 | Identifying narratives

The most challenging task the search tool has been used for is the attempt to automatically detect narratives in the interview corpus. What makes identifying narratives particularly difficult is that narrative is a function of language, not a form. Furthermore, narrative studies have not indicated particular linguistic forms present in every narrative. Thus far, computational narratology has aimed at detecting individual local narrative features from a text globally identified as narrative by humanities scholars. These features have included temporal sequences (Bögel, Strötgen, & Gertz, 2015) and certain types of action and actant roles (Droog-Hayes, Wiggins, & Purver, 2018; Ouyang & McKeown, 2014). Our model identifies complete narratives from a corpus not globally narrative in nature. Due to the nature of the interviews as retrospection of the MPs' careers, informants report and describe the past, which is why the linguistic surface between narrative and other types of reporting past events is of special interest in this project.

Narrative studies is a multi- and interdisciplinary field without shared understanding on how to define narrative. Three basic approaches include the cognitive, which defines narrative as a tool to make sense of the world, the rhetoric, which sees narrative as an intentional act of persuasion, and the semiotic, which emphasizes narrative as an articulation of some story content (see Hatavara & Toikkanen, 2019). Most definitions of narrative recognize at least two components in any narrative: story events and their organization into a narrative (Rimmon-

Kenan, 2006). Still, as noted by David Herman (2009, pp. 1–2), not all representations of a sequence of events are narratives. Our model provides a methodology to identify prototypical narratives and thus distinguish narratives from other representations of events. It seeks to detect narratives in natural language use, but—given the goals of the project and the complexity of defining a narrative—exact matching of the narratives' start and end points was not a priority.

Herman (2009, p. 14) identifies four elements of a prototypical narrative: (a) a narrative representation is situated, that is, it occurs in a specific occasion for telling, (b) the representation is about particularized events in a structured time-course, (c) the represented events introduce a disruption in the represented world, and (d) the representation conveys how it feels for a human-like agent to live through the represented events. Unfortunately, none of these basic elements have been linked to particular linguistic markers. In order to operationalize these elements of narrative for automatic detection, our model needs to (a) distinguish between the time points of the represented events and the situated telling about them, therefore detecting both situated telling and a represented time-course, (b) locate an experiencer in the represented events in order to recognize the portrayal of someone living through them.

We have a subcorpus of narratives that were manually annotated by experts in narrative studies, who made systematic interpretative decisions based on elements proposed by Herman. However, this subcorpus is not large enough to use data-intensive machine learning to identify narratives or to address the two needs of operationalization directly. Therefore, we utilize linguistic features, producing a rule-based method. Since a narrative incorporates a relation between two points of time and two subjectivities, we intend to locate constructions of two distinct elements. First, a sentence is required to

build the relation between the point of the telling and the point of the told—the nuclear sentence. Second, other sentences are needed to portray the events told—the context sentences. A relationship between a nuclear sentence and several context sentences is expected to bring together the situated telling and the experience in the represented time-course.

Thus, the model first searches for a nuclear sentence featuring any of the following: (a) a finite verb in the perfect tense; (b) a when-clause; (c) a third-person form of a speech act verb (see Pajunen 2001). For example, any sentence containing “has gone,” “have left,” or a third-person form of verbs such as “say,” “speak,” or “tell” would qualify as a nuclear sentence (see also Table 3 for a broader example). Second, the model looks ahead and behind each nuclear sentence for context sentences, which must contain a finite verb in preterite or pluperfect tense (in the active voice). Third, the model forms sequences from these two types of sentences: the nuclear sentence must be followed by at least two context sentences. When this condition is met, the sequence is extended by further context sentences as long as they directly precede or follow the nuclear sentence. We call the complete sequences found by the model passages.

In the implementation of this model, detecting when-clauses was fairly simple, given that this conjunction itself rarely appears without its clause. The positive preterite is morphologically distinct in Finnish and directly identified by the parser. However, negative finite forms as well as perfect and pluperfect tenses are periphrastic constructions (verb *olla*, to be, as an auxiliary verb), which makes their identification more complex. Due to Finnish language's grammatically free word order, these verb constructions can also split and the distance between their parts can be several words.

It was initially intended to use the parser's UD information, with the assumption that the auxiliary verb's

TABLE 3 Example of a manually annotated narrative found by the model

Minä en ollut AKS:n jäsen, enkä tuntenut silloinkaan vetoa IKL:ään	I was not a member of AKS, nor did I feel attracted to IKL, even then
Veljeni, joka sitten kaatui, hän oli enemmän kallellaan sinne IKL:ään päin	My brother, who was later killed in action, he was more prone to IKL
Mutta minä olin siinä suhteessa pidättyväisempi	But I was more reserved in that matter
Minä muistan hyvin, minulla oli opettajana sellainen [nimi1] -niminen, siihen aikaan jo vanha opettaja ja hän oli innokas IKL:läinen, niin kuin maalaiskansakoulun opettajat usein olivat	I remember it well, I had, as a teacher, this [name1], at the time already an old teacher, and he was an eager member of IKL, just as teachers in the countryside often were
Veljeni kertoi, että hän oli kerran sanonut minusta se [nimi1], he olivat kaksi IKL:läistä, että siinä [haastatellun nimi] on toinen henki, kun minä olin hiukan kriittisempi	My brother told me what he had once said about me, this [name1], they were two members of IKL, that this [the interviewee's name] has another kind of attitude, since I was a bit more critical

TABLE 4 Performance of the model on annotated data

Interview	Sentences	Annotated narratives	Annotated narratives found by the model	Passages found by the model	Passages not overlapping with any narratives
1	1,386	30	24 (80%)	46	21 (46%)
2	3,392	50	45 (90%)	131	81 (62%)
3	1,602	21	20 (95%)	41	23 (56%)
4	1,775	12	9 (75%)	64	52 (81%)
5	2,657	22	20 (91%)	95	73 (77%)
6	2,700	37	32 (86%)	86	53 (62%)
7	2,758	37	35 (95%)	98	55 (56%)

dependency would point to its participle. Thus, a combination of a participle in the nominative case and an auxiliary verb referring to it would indicate a perfect tense form. The auxiliary's form, much like in English, would then determine whether (indicative) perfect, pluperfect or perhaps conditional perfect is in question. Yet the parser's syntactic annotations were not as reliable as morphological ones: we noticed that the auxiliary verb was sometimes linked to a different word than its participle, and "backtracking" along the dependency chain from either word tended to introduce false positives.

A more practical approach that was ultimately discovered to identify perfect tense tends to focus only on the morphology, completely ignoring UD data. It extracts only the verb forms from a given sentence and analyzes them left-to-right, trying to bring together all the components of a potential compound form. When a verb is found that may act as an auxiliary, the algorithm looks ahead to see if it is indeed followed by a participle or is in fact a self-sufficient main verb. In the former case, the perfect tense is in question and its exact characteristics will be derived from the auxiliary. In the latter case, the parser's annotation of the main verb determines its voice, person, number, and tense.

Table 3 presents an example discovered by the model and manually annotated as a narrative, with the original text in the left column and the English translation in the right one.

The last sentence of Table 3 is the nuclear sentence with two speech act verbs (in bold) in third person (each referring to a different person). Besides this annotated narrative, the passage identified by the model also includes as many as 33 following sentences. They are in preterite but report varying happenings and are not interpretatively part of this narrative. The first three sentences preceding the nuclear sentence are also in preterite. They belong to the narrative as portrayals of the experience of the speaker and his brother living through the past

events. The endpoints of all the narratives were encoded in the annotation, so that the overlap between them and the model's output could be observed.

The model's accuracy was evaluated by running it on a subcorpus of seven manually annotated interviews and comparing the passages found by the model with the sentences belonging to actual narratives in the annotation. Table 4 illustrates the model's findings: a narrative was considered "found" if at least one sentence of it was retrieved by the model, and a single narrative's sentences could be split between multiple passages.

Comparing the model's findings with annotated material indicates that there is substantial overlap with actual narratives, but few precise matches. Many passages found by the model, like the example above, include long chains of past-tense sentences, going beyond the boundaries of the actual narrative. The rarity of exact matches is expected, since even expert annotators often disagree with each other on the beginnings and ends of narratives. Another significant issue is the presence of texts that should be considered mere instances of reporting, not of narration. The nature of speech act verbs allows them to be employed for reporting purposes ("they said," "he told") without establishing narrative structures. Regarding the portrayal of experience as a central feature of narratives, we expect verbs in the categories of mental action as well as affect and emotion also to be frequent in narratives besides speech act verbs. At this stage, the speech act verbs is the only category applied, but mental action verbs and verbs of affect may be examined in the future.

Recognizing narratives is a hard task, as even specialists often need to discuss whether a particular passage of text contains a narrative or not. Doing that computationally is apparently also hard. In our case, of course, we have a special domain and a specific dataset, which may on the one hand make the task easier in some aspects while on the other hand make it more difficult in others. In such a setting, analyzing the quality of our results is hard as there is no real reference point.

Let us consider, as a point of comparison, a classic hard read comprehension task where the problem consists of a passage—a news article, and a task of guessing which entity appearing in the news article belongs to a position in the article's bullet point (Chen, Bolton, & Manning, 2016). The guess is performed from the entities appearing in the news article text and the entity to be guessed is replaced by a placeholder. In this relatively hard novel task, the first attempts reached up to around 56% accuracy without machine learning and using neural networks, and nearly 70% accuracy using neural networks. Later the neural solutions have reached around 80% accuracy. In our case, neural networks were not applicable due to a small amount of annotated data. Surely, our problem is rather different but generally we see that we have a reasonable solution for a genuinely hard NLP problem. Our model found a large fraction (from 75 to 95%) of the existing narratives, however it still suffers from a relatively high number of false positives.

4 | CONCLUSIONS

We developed two corpora and a search system to facilitate utilization of Natural Language Processing in digital humanities research. The system was used in three case studies demonstrating three different ways of using grammatically parsed corpora to make sense of a large amount of parliamentary data. First, in the study of power, keyword searches were utilized to find relevant material for close reading in order to answer qualitative research questions. Second, in the case of isms, metadata integrated into the corpora (date and party-political affiliation) and a user-demanded feature of the search system (flexible window search) enabled humanities scholars to generate several hypotheses on the major changes in the Finnish political culture. Third, in the analysis of narratives, the search system facilitated experts of narrative studies and computer scientists to develop a computational method to identify a complex concept in the dataset. Our system had a key role in successful treatment of the humanities research tasks.

Thus, instead of trying to define what the digital humanities should be, this article has shown through concrete examples that the distinction between contemporary humanities and computational humanities is not always that useful. Inquiry can seldom be reduced to computational modeling of humanities research questions. Instead, studies using digital resources relate to many different needs in the context of humanities: not only modeling, but also finding information more quickly and effectively and generating hypotheses with the help of simple quantifications. Even in complex modeling tasks, such as our case

on identifying narratives, there was a need to revert back to simpler methods in order to facilitate contextual knowledge that supports interpretation of models. From an information science perspective, it is essential that a tool facilitates exploration and gradual development of high-quality corpus utilization. However, also from the perspective of computational humanities the interpretative element is strong. Even if modeling is successful, the most crucial part of a study lies in the interpretation of the results and that cannot be performed without the contemporary humanities perspective.

The parsed corpora together is a collection of both broad size, with over 6 million sentences and extendable in principle to about 15 million using older plenary sessions, and superb quality. Documents belonging to the collection have been proofread and refined repeatedly, to the point where spelling errors are virtually nonexistent. Even spoken language, which differs noticeably from the literary standard in Finland, was not an issue: parliamentary speech has much in common with written language, and the presence of spoken forms is minimal since the transcription process generally converts them to standard language (Voutilainen, 2017).

Directions to further develop the system fall into three categories: (a) extension and further cleaning of the data in the corpora, (b) improving the linguistic processing of the data, and (c) improving the search system.

While the interview data will only grow as new interviews are hopefully produced in the future, there are still parliamentary records from the past that are only available in pdf format, having several technical complications for utilization, as the quality is variable and text placement in the pages is not optimal for OCR. There are separate attempts to produce cleaned-up versions reaching further in the past. The other option is to invest manual work to improve the quality resulting from OCR. Utilization of the pdfs is risky both considering the actual texts and the metadata. As for the pdf-format data, we have been able to correct metadata errors gradually by adjusting the post-OCR parsing process and trying to account for various edge cases.

The improvement of linguistic processing relates to the utilization of UD parsed data. While the parser mostly produces the necessary useful information, we have had to post-process some of the UD information. For instance, sometimes adjectives point to unrelated words, so they are checked for agreement in their case and number, or lack of case and number altogether. Multiple adjectives for a noun are sometimes arranged in the UD in such a way that a more distant adjective points to another less distant adjective, not directly to the noun, which also requires post-processing. Parts of compound verb forms do not necessarily retain UD links to each

other, either, partly due to the freedom of Finnish word order and the possibility to split verb phrases. Using both morphological and syntactic annotations may be required to detect such compounds more accurately.

As for the actual search system, the planned further developments are related to automatically producing more statistics, and exporting models for further analysis using, for example, a data analysis/statistical software.

A particular strength of our search system compared to for instance the Korp system used by the Finnish Language Bank (Kielipankki) is that it is relatively easy to use the context around the searched term when searching the corpora. Based on the queries performed and requested by the users, more statistics for co-occurrences would be useful as search criteria as well as search output. Also, co-occurrences and interactions of words, temporal expressions, and metadata are of interest to the users. This is already supported to some extent in the search system, but we plan to automate tests of statistical significance as a part of our future work.

We created a system for the corpora with its own search mechanism, and utilized the system with the contents of Plenary sessions of Finnish Parliament to analyze meaningful research problems in humanities. The volume of material we are dealing with would not be feasible to study manually. The grammatical information played a central role in the utilization of the system for digital humanities.

ORCID


Jani Marjanen  <https://orcid.org/0000-0002-3085-4862>

Jussi Kurunmäki  <https://orcid.org/0000-0002-8089-8468>

Timo Nummenmaa  <https://orcid.org/0000-0002-9896-0338>

Matti Hyvärinen  <https://orcid.org/0000-0002-1145-9656>

Jaakko Peltonen  <https://orcid.org/0000-0003-3485-8585>

Jyrki Nummenmaa  <https://orcid.org/0000-0002-7476-7840>

REFERENCES

- Ahonen, P. (2015). Institutionalizing Big Data methods in social and political research. *Big Data & Society*, 2(2), 205395171559122. <https://doi.org/10.1177/2053951715591224>
- Ahonen, P. (2018). Laskennalliset koneoppimisen menetelmät politiikan tutkimuksen kannalta: Institutionaalinen tarkastelu ja tieteenfilosofisen ja teoreettisen syventämisen tarve. *Politiikka*, 60(2), 157–163.
- Ahonen, P., & Wiberg, M. (2018). Laskennallisten menetelmien mahdollisuuksia politiikan tutkimuksessa. *Politiikka*, 60(1), 38–46.
- Biemann, C., Crane, G. R., Fellbaum, C. D., & Mehler, A. (2014). Computational humanities—Bridging the gap between computer science and digital humanities (Dagstuhl Seminar 14301). *Dagstuhl Reports*, 4(7), 80–111. <https://doi.org/10.4230/DAGREP.4.7.80>
- Bögel, T., Strötgen, J., & Gertz, M. (2015). A hybrid approach to extract temporal signals from narratives. In *Proceedings of German Society for Computational Linguistics and Language Technology '15*.
- Bonin, H. (2020). From antagonist to protagonist: “Democracy” and “people” in British parliamentary debates, 1775–1885. *Digital Scholarship in the Humanities*, 35(4), 759–775. <https://doi.org/10.1093/llc/fqz082>
- Chen, D., Bolton, J., & Manning, C. D. (2016). A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of ACL 2016*. Association for Computational Linguistics.
- Crum, W. B., Angello, A., Liu, X., & Campion, C. (2019). Enabling interdisciplinary instruction in computer science and humanities: An innovative teaching and learning model customized for small Liberal arts colleges. In J. M. F. Rodrigues, P. J. S. Cardoso, J. Monteiro, R. Lam, V. V. Krzhizhanovskaya, M. H. Lees, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational science – ICCS 2019* (Vol. 11540, pp. 389–400). Springer. https://doi.org/10.1007/978-3-030-22750-0_31
- Curran, B., Higham, K., Ortiz, E., Vasques, & Filho, D. (2018). Look who’s talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches. *PLoS ONE*, 13(6), e0199072. <https://doi.org/10.1371/journal.pone.0199072>
- Droog-Hayes, M., Wiggins, G., & Purver, M. (2018). Automatic detection of narrative structure for high-level story representation. In *Fifth AISB symposium on computational creativity*. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.
- Freeden, M. (2017). After the Brexit referendum: Revisiting populism as an ideology. *Journal of Political Ideologies*, 22(1), 1–11. <https://doi.org/10.1080/13569317.2016.1260813>
- Grimmer, J. (2013). *Representational style in congress: What legislators say and why it matters*. Cambridge University Press.
- Hatavara, M., & Toikkanen, J. (2019). Sameness and difference in narrative modes and narrative sense making: The case of Ramsey Campbell’s “The Scar”. *Frontiers of Narrative Study*, 5(1), 130–146.
- Haugaard, M., & Clegg, S. R. (2009). Introduction: Why power is the central concept of the social sciences. In S. R. Clegg & M. Haugaard (Eds.), *The SAGE handbook of power*. SAGE.
- Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., & Ginter, F. (2014). Building the essential resources for Finnish: The Turku dependency treebank. *Language Resources and Evaluation*, 48(3), 1–39.
- Herman, D. (2009). *Basic elements of narrative*. Wiley-Blackwell.
- Holstein, J. A., & Gubrium, J. F. (1995). *The active interview*. SAGE.
- Höpfl, H. M. (1983). Isms. *British Journal of Political Science*, 13(1), 1–17. <https://doi.org/10.1017/S0007123400003112>
- Hyvärinen, M. (2003). Valta. In *Käsitteet liikkeessä: Suomen poliittisen kulttuurin käsitehistoria*. Vastapaino.
- Ihalainen, P., Ilie, C., & Palonen, K. (Eds.). (2016). *Parliament and parliamentarism: A comparative history of a European concept*. Berghahn Books.
- Ihalainen, P., & Palonen, K. (2009). Parliamentary sources in the comparative study of conceptual history: Methodological aspects and illustrations of a research proposal. *Parliaments, Estates and Representation*, 29(1), 17–34. <https://doi.org/10.1080/02606755.2009.9522293>

- Ilie, C. (Ed.) (2010). European parliaments under scrutiny. In *Discourse strategies and interaction practices*. John Benjamins.
- Kettunen, K., & La Mela, M. (2020). Digging deeper into the Finnish parliamentary protocols: Using a lexical semantic tagger for studying meaning change of everyman's rights (allemanräätten). In S. Reinsone, I. Skadina, A. Baklāne, & J. Daugavietis (Eds.), *DHN 2020: Digital humanities in the Nordic countries. Proceedings of the digital humanities in the Nordic countries fifth conference (CEUR workshop proceedings; no. 2612)* (pp. 63–80). Retrieved from CEUR Workshop Proceedings. <http://ceur-ws.org/Vol-2612/paper5.pdf>
- Koselleck, R. (2011). Introduction and prefaces to the geschichtliche grundbegriffe. *Contributions to the History of Concepts*, 6(1), 1–37. <https://doi.org/10.3167/choc.2011.060102>
- Kurunmäki, J., & Marjanen, J. (2018a). A rhetorical view of isms: An introduction. *Journal of Political Ideologies*, 23(3), 241–255. <https://doi.org/10.1080/13569317.2018.1502939>
- Kurunmäki, J., & Marjanen, J. (2018b). Isms, ideologies and setting the agenda for public debate. *Journal of Political Ideologies*, 23(3), 256–282. <https://doi.org/10.1080/13569317.2018.1502941>
- Kurunmäki, J., Nevers, J., & te Velde, H. (Eds.). (2018). *Democracy in modern Europe: A conceptual history*. Berghahn Books.
- Loukasmäki, P., & Makkonen, K. (2019). Eduskunnan täysistunnon puheenaiheet 1999–2014: miten käsitellä LDA-aihemalleja? *Politiikka*, 61(2), 127–159.
- Lytard, J.-F. (1979). *La condition postmoderne: Rapport sur le savoir*. Éditions de Minuit.
- Mäkelä, E., Tolonen, M., Marjanen, J., Kanner, A., Vaara, V., & Lahti, L. (2019). Interdisciplinary collaboration in studying newspaper materiality. In S. Krauwer & D. Fišer (Eds.), *Twin Talks workshop at DHN 2019* (pp. 55–66). Retrieved from CEUR Workshop Proceedings. <https://cst.dk/DHN2019/DHN2019.html>
- Nelimarkka, M. (2019). Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: kriittisiä havaintoja. *Politiikka*, 61(1), 6–33.
- Ouyang, J., & McKeown, K. (2014). Towards automatic detection of narrative structure. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. European Language Resources Association (ELRA).
- Pajunen, A. (2001). *Argumenttirakenne: asiointilojen luokitus ja verbien käyttäytyminen suomen kielessä*. SKS.
- Piotrowski, M. (2019). Accepting and modeling uncertainty. In *Die Modellierung Des Zweifels – Schlüsselideen Und -Konzepte Zur Graphbasierten Modellierung von Unsicherheiten*. ZfdG - Zeitschrift für digitale Geisteswissenschaften. https://doi.org/10.17175/SB004_006A
- Piotrowski, M. (2020). *Ain't no way around it: Why we need to be clear about what we mean by "digital humanities"*. <https://doi.org/10.31235/osf.io/d2kb6>
- Plenary Sessions of the Parliament of Finland. (2020). Kielipankki Korp Version 1.1. Transcriptions of the plenary sessions of the parliament of Finland from 10.09.2008 to 1.7.2016, is available in Kielipankki, the Language Bank of Finland (through the Korp service). Retrieved from <http://urn.fi/urn:nbn:fi:lb-2017020202>
- Rimmon-Kenan, S. (2006). Concepts of narrative. In M. Hyvärinen, A. Korhonen, & J. Mykkänen (Eds.), *The travelling concept of narrative (studies across disciplines in the humanities and social sciences 1)* (pp. 10–19). Helsinki Collegium for Advanced Studies.
- VoDe Corpora. (2020). Parliamentary text corpus. Voices of Democracy Project, Plenary sessions of the parliament of Finland from 1980 to 2018. Interview corpus. Voices of Democracy Project, 404 veteran parliamentarians' interviews. Both corpora grammatically parsed. Requires separate user rights. Available at <https://vode.uta.fi>.
- Voutilainen, E. R. J. (2017). The regulation of linguistic quality in the official speech-to-text reports of the Finnish parliament. *CoMe: Studies on Communication and Linguistic and Cultural Mediation*, 61–73. <http://comejournal.com/home-en/>.

How to cite this article: Andrushchenko, M., Sandberg, K., Turunen, R., Marjanen, J., Hatavara, M., Kurunmäki, J., Nummenmaa, T., Hyvärinen, M., Teräs, K., Peltonen, J., & Nummenmaa, J. (2021). Using parsed and annotated corpora to analyze parliamentarians' talk in Finland. *Journal of the Association for Information Science and Technology*, 1–15. <https://doi.org/10.1002/asi.24500>