Suomi24 Corpus Analysis

1st Amirhossein Ghaffari
Student ID: Haven't got yet
University of Oulu
Oulu, Finland
amirhossein.ghaffari@oulu.fi

2nd Fatemeh Mahjouyanmoghaddam
Student ID:2208093
University of Oulu
Oulu, Finland
fatemeh.mahjouyanmoghaddam@student.oulu.fi

3rd Ghazal Vatankhahan
Student ID:2209705
University of Oulu
Oulu, Finland
ghazal.vatankhahan@student.oulu.fi

***Abstract*—Linguistic patterns in the Finnish Suomi24 corpus were investigates from 2001 to 2020. Employing Zipf's and Heaps' laws, the research focuses on the occurrences of a keyword, the evolution of vocabulary, lexical distance analysis, sentiment evaluation, and topic modeling, with a particular emphasis on the Finnish translation of the term "climate change". The methodology involved dissecting discussions to trace shifts in topics over the years, thereby uncovering the dynamic nature of public discourse within the corpus. The findings reveal significant insights into the linguistic patterns, sentiment fluctuations, and thematic trends prevalent in the corpus, offering a nuanced understanding of the evolving public perception and discourse on climate change. This analysis not only sheds light on the linguistic characteristics of the Suomi24 corpus but also provides a detailed overview of societal attitudes and ideas surrounding "climate change", as reflected in public discussions over two decades.**

***Index Terms*—NLP, Zipf's Law Analysis, Heaps' law, Finnish Corpus Study, Soumi24, Linguistic Pattern Exploration, Temporal Keyword Occurrence, Sentiment analysis**

I. Introduction

The Suomi24 forum is one of Finland's largest online discussion platforms that reflects people's opinions in Finnish society. The forum contains several different main topics such as hobbies, society, economy, and health. Each topic has some sub topics. All the main topics, categories, subtopics, subcategories are provided by Suomi24 and users are not able to add new or edit existings [1]. It provides valuable data for any societal analysis. It can be used in diverse applications.

Corpus linguistics provides a useful platform for analyzing linguistic patterns and textual dynamics across multiple languages and situations. This study investigates the huge Suomi24 corpus, a large Finnish text dataset containing debates from 2001 to 2020. Our goal is to unravel the complicated linguistic fabric inside this corpus to understand how language changes over time and interacts with discourse themes. A powerful technique for revealing linguistic structures, monitoring the use of keywords, and evaluating the attitudes and themes that surface in conversations is corpus analysis. We investigate the use of the well-known Zipf's and Heaps' laws from the corpus linguistics field on the Suomi24 corpus.

The basic linguistic phenomenon known as Zipf's Law states that word frequencies in natural language have a power-law distribution [2]. Zipf's Law states that the frequency of a word's rank is as follows:

$$f(r) = \frac{1}{r^{\alpha}} \tag{1}$$

This is inversely related to its frequency, with α usually being near to 1 [2].

Heaps' Law is an empirical law describing the average growth in the number of unique elements (or records) when elements are randomly drawn without replacement from a statistical distribution. Specifically, in the context of word occurrences in natural language, Heaps' Law predicts the vocabulary size of a document based on its text size, which is the number of words it contains. This law states that the number of unique elements grows as $\alpha\kappa^{\beta}$ where α and β are application-dependent constants and $0<\beta<1$ with κ representing the number of drawings. The formula for Heaps' Law is [3]:

$$N(k) = \alpha\kappa^{\beta} \tag{2}$$

$N(k)$ represents the number of unique elements (vocabulary size). k is the number of drawings. α and β are constants that depend on the specific application.

This research project includes activities like keyword frequency tracking, sentiment analysis, and topic modeling. With these initiatives, we hope to shed light on the corpus's temporal language patterns, emotion shifts, and theme trends impacting debates throughout time. The source code of the project and processed data is available on the project GitHub repository [1].

II. Related work

Considering the fact that people talk about anything that concerns their life in the online forums, the discussions can be helpful to be used in many applications. For mining security-related comments in the Suomi24 online forum, a social network-based approach was presented. They collected a dictionary of keywords connected to Finland's national security using a student survey questionnaire. The keywords are then mapped to the Suomi24 corpus to create a social network that quantifies the dependency between the various vocabulary phrases. They have developed a tool that may be used to assess how strongly various terms are connected in Suomi24 conversations [4].

[1] https://github.com/Ahghaffari/suomi24_analysis

Another study that used this dataset, suggests a new approach to identifying hate speech in Finnish social media posts by combining a Finnish language model called FinBERT with a convolutional neural network (CNN). While FinBERT is used to extract semantic features from the text, CNN is used to extract features from the social media post's text. To determine whether a post contains hate speech or not, the features retrieved by CNN and FinBERT are merged and fed into a machine-learning classifier. Using a dataset of 10,000 social media posts from Finland that had been manually flagged for hate speech, they tested this approach. The findings showed their approach detected hate speech with a 95% accuracy rate [5].

A novel approach that synergizes multiple natural language processing tools specifically adapted for the intricacies of the Finnish language was presented in [6]. These methods are used to monitor disease-related conversations in the Suomi24 forum and extract health-related insights, providing a useful resource for medical professionals looking for novel approaches to patient care.

The dataset can be used in assessing depression and analysing the data as in [7] or exploring health aspects and health issues as in [8]. In the realm of health-related discussions on Suomi24, there is a significant study focused on cancer health forums. This research delved into the psychosocial factors prevalent in patient communities. Its innovative approach, combining ontology-based thread identification and co-occurrence analysis, revealed a strong emphasis on social factors, aligning with findings from clinical studies. This work not only underscores the potential of integrating online forum data with electronic medical records but also highlights the challenges in adapting NLP tools for Finnish language, a consideration pivotal to our study of linguistic patterns in the Suomi24 corpus. [9]

Understanding the linguistic nuances and textual dynamics in a large corpus helps us understand how social, cultural, and contextual factors change over time. The results can also guide the creation of information retrieval, sentiment analysis, and language technologies.

We will discuss the methodologies we employed, the tasks we completed, and the results we acquired while examining the Suomi24 corpus in the next sections of this report. We anticipate that this effort will enhance the science of corpus linguistics and knowledge of linguistic evolution in the Finnish language and other languages.

III. Methodology

A. *Data description and data parsing*

In this research the dataset which was used, was created by the data collection from Suomi24 platform. The dataset consists two parts, first part consists all the data from 2001 to 2017 [10] and the other one has the data from 2018 to 2020 [11]. In total there is 20 years of data available. The dataset consists of threads posts with their titles, comments, and so many features for each post like the authors name, datetime, topic name, and etc. The structure of the dataset is like YAML structure, but it is not a standard structure, So it needs a parser to parse and prepare the data.

The dataset is huge. In order to parse the data efficiently zip data parsed without extracting. Also, to process the data more resource efficiently threading used to distribute the tasks to many threads to do it in parallel. The dataset parsed line by line, the required features extracted from the dataset and saved for the furthur process. Only the main threads took into account for analysis as it is required in the project description.

B. *Data Preprocessing and Keyword Filtering*

In the first step of preprocessing the numerical values, URL links, emojis, and all the punctuations were removed from the dataset. The methodology was the NLTK library for text processing, including tokenization and stemming, and employed a SnowballStemmer specific to the Finnish language to reduce words to their base forms. The objective of this method is to filter text data from a huge dataset, which represents threads in Soumi24, and extract threads that contain a predefined keyword, "ilmastonmuutos" (climate change). This method is particularly useful for isolating discussions and content related to environmental topics, specifically climate change. The output of this stage is a dataset which preprocessed and it contains the threads that contain the required keyword.

C. *Vocabulary Calculation*

The method lies in the calculation of vocabulary size and total tokens for each year. This is achieved by employing the CountVectorizer from the scikit-learn library, which is used to transform the text data into a numerical format suitable for analysis. For each year, the total unique vocabulary size and the total number of tokens in the corresponding text data were calculated.

D. *Data Visualization*

Matplotlib library is used to create the plots. For showing the evolution of vocabulary size on a yearly basis, the evolution of vocabulary size over the years using a bar plot was visualized. This allowed us to observe trends and patterns in vocabulary changes within the Suomi24 corpus from 2001 to 2020.

E. *Assessing Vocabulary Growth and Heaps' Law Fit Over Time*

Heaps' Law and Linear Regression: Heaps' Law, as it was explained previously, was applied to model vocabulary growth. Linear regression was utilized to fit a model to the relationship between the logarithm of the vocabulary size and the logarithm of the number of tokens.

It was investigated whether a parametric model, specifically Heaps law, could be fitted to describe this relationship. The process involved the utilization of confidence values to establish the upper and lower bounds of the curve fitting and to evaluate the goodness of fit. Our approach is inspired by established statistical methods such as spacy.stats.linregress.

Confidence intervals are computed to estimate the range within which the vocabulary size is expected to fall outside

the calculated upper and lower bounds for each case. Different confidence levels (80%, 85%, 90%, and 95%) were considered, and prediction intervals were determined based on statistical analysis.

F. Analyzing Co-Occurring Words for Selected Keywords (2001-2020)

Identifying the ten most frequent co-occurring words for selected keywords across each year (2001-2020) was developed. The methodology involved tokenizing the discussion data, locating instances of the keywords, and extracting words within a three-word proximity. The context window is set to three units, and it captures the words surrounding the keyword in the text. A frequency distribution was used to tally these neighboring words, excluding the target keyword. It utilizes the NLTK library's FreqDist to compute word frequencies and ranks the top co-occurring words. These findings were visually represented through scatter plots, demonstrating co-occurring word patterns.

G. Visualization and Analysis of Temporal Evolution of Co-occurring Words

The method focuses on examining and visually representing how the frequency of co-occurring words changes over time. A pivotal step in the method involves pivoting the DataFrame to reorganize the data for visualization purposes. A heatmap was generated to illustrate the temporal evolution of the ten most frequent co-occurring words. The heatmap provides a visual representation of word frequency changes over the years, with colors indicating the relative frequency levels.

H. Scoring and Ranking of Discussion Threads

A systematic approach was offered to evaluate and rank discussion threads over multiple years, focusing on the discussion content's token count. The discussion score was determined by counting the number of tokens in the discussion part of each thread. This step ensures that the process can evaluate and rank the threads based on their textual content.

I. Temporal Topic Modeling and Analysis of Textual Data

For the implementation of topic modeling, Latent Dirichlet Allocation (LDA) was applied to textual data over a range of years, with a primary focus on extracting and examining topics and their associated keywords. For each year, the method organizes documents, separates text data, and constructs a dictionary to represent the unique terms within the text data. A corpus is also formed by converting the documents into a bag-of-words representation.

J. Sentiment-Enhanced Temporal Topic Modeling and Analysis of Textual Data

To sentiment Analysis, it was created a sentiment analyzer for the Finnish language using the AFINN lexicon. Each text, comprised of the thread title and content, was assigned a sentiment label (positive or negative) based on its sentiment score. Topic modeling was performed separately for positive and negative sentiments. For each year and sentiment category, an LDA topic model with two topics was created. The most representative keywords for each topic were extracted. One thing that needs mentioning is that the AFINN package which is installed by package installers like pip doesn't support the finnish language so in order to use that we needed to utilize the latest update on its github page.

K. The state of the art method

To use a state of the art method to achieve the specifications we could use Non-Negative Matrix Factorization(NMF) [12], [13] and also FinBERT [14] available in which are the two state of the art methods and they can be used in the same way as the LDA method. The NMF method selected to be used in the corpus topic modeling and the topic of the threads for each year extracted like the other method. Two topics were extracted for each years corpus.

IV. Results and Discussion

A. Evolution of Vocabulary Size in Suomi24 Threads

We extracted the vocabulary size and total tokens for each year. Interestingly, in 2001, the dataset did not contain any keywords. This outcome is shown for the years 2001–2020 in "Fig. 1".

```
Year 2001: Vocabulary Size = 0, Total Tokens = 0
Year 2002: Vocabulary Size = 78, Total Tokens = 108
Year 2003: Vocabulary Size = 745, Total Tokens = 1219
Year 2004: Vocabulary Size = 2853, Total Tokens = 7675
Year 2005: Vocabulary Size = 4562, Total Tokens = 10283
Year 2006: Vocabulary Size = 15314, Total Tokens = 53768
Year 2007: Vocabulary Size = 17824, Total Tokens = 66641
Year 2008: Vocabulary Size = 17039, Total Tokens = 62184
Year 2009: Vocabulary Size = 13873, Total Tokens = 43210
Year 2010: Vocabulary Size = 11752, Total Tokens = 36011
Year 2011: Vocabulary Size = 8955, Total Tokens = 23165
Year 2012: Vocabulary Size = 5867, Total Tokens = 12469
Year 2013: Vocabulary Size = 5838, Total Tokens = 14827
Year 2014: Vocabulary Size = 5855, Total Tokens = 14470
Year 2015: Vocabulary Size = 6528, Total Tokens = 14243
Year 2016: Vocabulary Size = 6472, Total Tokens = 16788
Year 2017: Vocabulary Size = 5660, Total Tokens = 12339
Year 2018: Vocabulary Size = 8393, Total Tokens = 22606
Year 2019: Vocabulary Size = 13039, Total Tokens = 43129
Year 2020: Vocabulary Size = 8536, Total Tokens = 25495
```

Fig. 1. Evolution of vocabulary size

Subsequently, as the years progress from 2001 to 2007, the vocabulary size consistently increases. This suggests that the text data becomes more diverse, containing a larger variety of unique terms. After 2007 we have fluctuations in the number of vocabulary size and the number of tokens. Additionally, "Fig. 2" depicts a graph of the evolution of vocabulary size over time.

B. Heaps' Law Assessment and Confidence Interval Analysis for Vocabulary Growth Dynamics

In the absence of data containing the specified keyword in the year 2001, a singular post is observed in the year 2002, and a mere two posts are available for analysis in the year 2003. Consequently, the dataset utilized for all subsequent

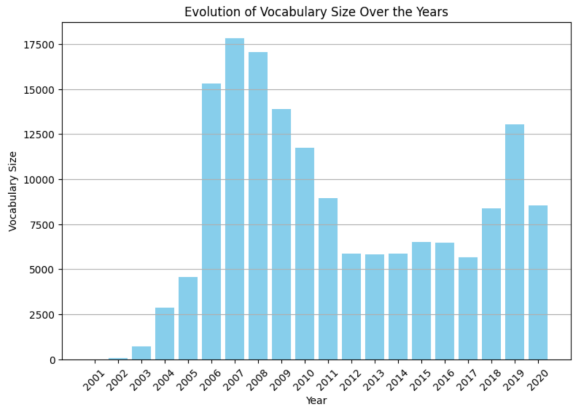

Fig. 2. The graphical evolution of vocabulary size

analyses spans the years 2004 to 2020, as these years provide a more robust and representative corpus for comprehensive investigation.

The following plots showcase the relationship between vocabulary size and the number of tokens for each year, providing valuable insights into the evolution of language and information complexity in the Suomi24 corpus from 2001 to 2020. The blue line represents the actual vocabulary size observed in the dataset, while the orange dashed line signifies the predicted vocabulary size based on Heaps' law. The shaded regions around the predicted line illustrate the confidence bounds, which help quantify the uncertainty in the predictions. By examining these plots for each year, we can discern how the corpus's vocabulary grows over time and assess the goodness of fit for the Heaps' law model. The analysis will showcase the graphical representation of vocabulary growth for the years 2004 and 2020 "Fig. 3" and "Fig. 4". The inclusion of additional graphs for the intervening years from 2005 to 2019 is deferred to the appendix.

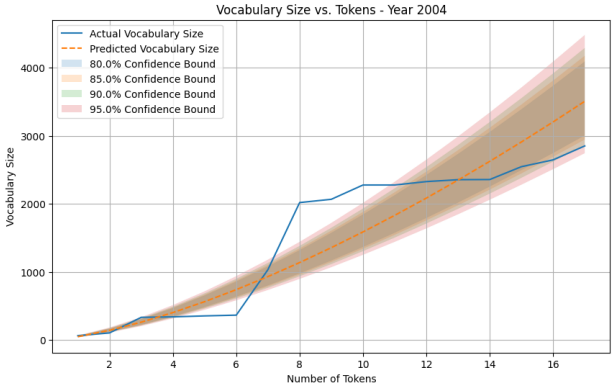

Fig. 3. Vocabulary Growth and Heaps' Law Fit in the Year 2004

In the context of the year 2004, the statistical analysis reveals noteworthy observations based on varying confidence levels. At an 80.0% confidence level, 12 data points lie outside the established bounds, indicative of a broader variability within the dataset. Similarly, at confidence levels of 85.0%, 90.0%, and 95.0%, the number of points outside bounds decreases progressively, with 11, 10, and 6 data points, respectively.

The percentage confidence levels in the plots—80.0%, 85.0%, 90.0%, and 95.0%—represent the statistical certainty of the Heaps law model-derived vocabulary size forecasts. Higher degrees of confidence correspond to narrower prediction intervals, indicating greater confidence in the model's correctness. "80.0% Confidence," for example, denotes a bigger prediction range with more potential outliers, but "95.0% Confidence" denotes a tighter range with greater confidence in the model's precision. These ratings aid in determining the dependability and confidence in the vocabulary size projections.

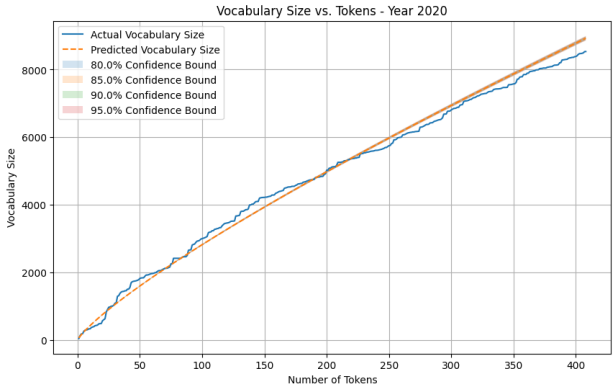

Fig. 4. Vocabulary Growth and Heaps' Law Fit in the Year 2020

In the context of the year 2020, at an 80.0% confidence level, a substantial 389 data points extend beyond the established bounds, reflecting a considerable dispersion within the dataset. This trend persists, albeit with a diminishing number of outliers, as the confidence level increases to 85.0%, 90.0%, and 95.0%, resulting in 385, 381, and 378 points outside bounds, respectively.

C. Top Ten Co-Occurring Words for Selected Keywords (2001-2020)

We extracted the ten most frequent words co-occurring with keywords within a 3-word span each year. "Fig. 5" provides a visual representation of how these co-occurring words change over time.

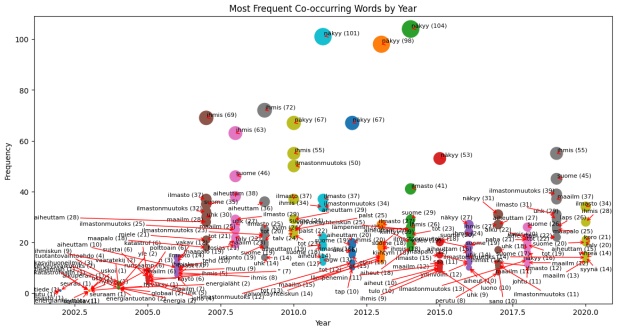

Fig. 5. Scatter Plots Illustrating Co-Occurring Word Patterns

D. *Annual Changes in the Frequencies of the Top 10 Co-occurring Words*

Another plot ("Fig. 6") allows us to observe the annual changes in the frequencies of the 10 most frequent co-occurring words. Each row corresponds to a year, each column represents a word, and the color intensity reflects word frequency. Darker colors indicate higher usage, while lighter shades or empty cells suggest reduced or no usage for that word in that particular year.

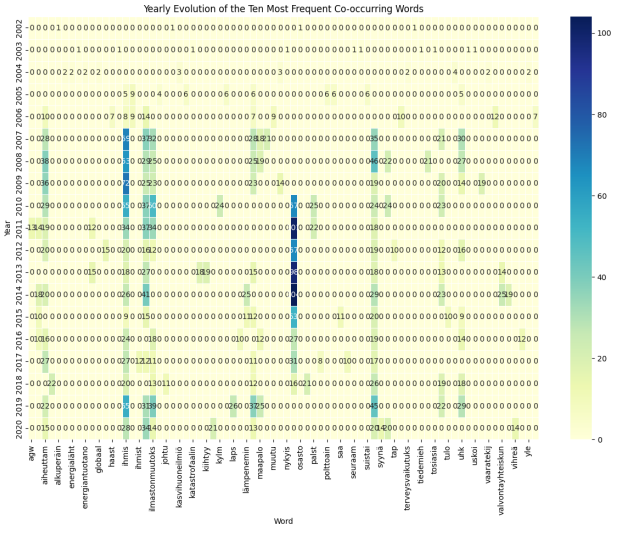

Fig. 6. Annual Variation in Frequencies of the Top 10 Co-occurring Words

E. *Quantitative Evaluation of Discussion Threads*

We assessed discussion threads from various years, ranking them based on the discussion score. The top 5 threads for each year are then saved in a file named **top_ranked_threads_and_perceptions.txt** on our GitHub; the file includes the year, discussion score, thread title, and thread text. We can see the outputs for 2020 in the "Fig. 7" as an example. The full output can be found in the GitHub repository.

Discussion scores 2020

Discussion score	Thread title
396	ilmastonmuutos huijaus globalisaatio nwo n välin
390	puk poro lähten liik etel
385	greta thurnberg ilmasto lap
346	poliittin ilmastonmuutos
343	valtiosihteer risto volon arvostel timo soin ulkoministerikaut

Fig. 7. Scoring and Ranking Based on Token Count

The analysis discerns variations in the number of tokens within thread texts, noting instances where a substantial token count does not necessarily correlate with relevance to the specified keyword. While some threads exhibit extensive token content that is unrelated to the keyword and lacks contextual relevance in the discussion, others with comparatively fewer tokens maintain a more substantive connection to the keyword. This observation challenges the simplistic assumption that the sheer number of tokens inherently influences the relevance of a thread to the specified keyword, emphasizing the need for nuanced evaluation criteria in determining topical alignment within the discussions.

F. *Latent Dirichlet Allocation (LDA) Analysis on Textual Data Across Multiple Years*

This output represents the result of a topic modeling analysis conducted on threads and their related discussions for each year, grouped separately. The objective of this analysis was to identify prevalent topics within these discussions for every year, using Latent Dirichlet Allocation (LDA) with two topics and five keywords per topic. The results are organized in a tabular format, showcasing the key topics and associated keywords for each year. "Fig. 8" illustrates the first topics and "Fig. 9" shows the second topics.

Year	Topic 1 Keywords
2004	0.088*"suome" + 0.059*"ilmastonmuutoks" + 0.056*"energia" + 0.050*"alue" + 0.041*"käytö"
2005	0.089*"ydinvoim" + 0.059*"the" + 0.048*"käytö" + 0.041*"raport" + 0.022*"climat"
2006	0.013*"talo" + 0.012*"niinistö" + 0.012*"energia" + 0.011*"suome" + 0.007*"tuot"
2007	0.012*"ilmasto" + 0.011*"ilmastonmuutoks" + 0.010*"the" + 0.007*"ihmis" + 0.006*"auto"
2008	0.011*"ihmis" + 0.010*"suome" + 0.010*"maailm" + 0.006*"osa" + 0.006*"ilmastonmuutoks"
2009	0.019*"ilmasto" + 0.019*"ihmis" + 0.013*"ilmastonmuutoks" + 0.009*"maapalo" + 0.006*"maailm"
2010	0.017*"the" + 0.014*"maailm" + 0.011*"suome" + 0.011*"maapalo" + 0.010*"ihmis"
2011	0.033*"ilmasto" + 0.023*"näkyy" + 0.022*"the" + 0.018*"ilmastonmuutoks" + 0.015*"of"
2012	0.032*"ilmastonmuutoks" + 0.023*"ihmis" + 0.017*"maapalo" + 0.015*"maailm" + 0.015*"ilmasto"
2013	0.020*"maailm" + 0.019*"lopu" + 0.016*"ilmastonmuutoks" + 0.011*"ihmis" + 0.010*"työpaik"
2014	0.033*"ilmasto" + 0.025*"näkyy" + 0.021*"ilmastonmuutoks" + 0.020*"suome" + 0.018*"ihmis"
2015	0.033*"ilmastonmuutoks" + 0.027*"ilmasto" + 0.026*"näkyy" + 0.014*"lämpenemin" + 0.013*"tulo"
2016	0.033*"suome" + 0.022*"maailm" + 0.020*"ilmastonmuutoks" + 0.016*"venäj" + 0.016*"ihmis"
2017	0.026*"ilmasto" + 0.025*"trump" + 0.025*"suome" + 0.023*"ydinvoim" + 0.022*"ilmastonmuutoks"
2018	0.020*"suome" + 0.018*"ilmasto" + 0.016*"ihmis" + 0.015*"ilmastonmuutoks" + 0.010*"maailm"
2019	0.014*"ilmastonmuutoks" + 0.012*"ihmis" + 0.010*"puolue" + 0.009*"suome" + 0.009*"vihreä"
2020	0.035*"kenkäpar" + 0.022*"keng" + 0.020*"kuoma" + 0.018*"talv" + 0.016*"suome"

Fig. 8. Topic Modeling Analysis Results, Topic1 keywords

Year	Topic 2 Keywords
2004	0.073*"maailm" + 0.072*"olem" + 0.049*"kasvihuoneilmiö" + 0.047*"kehity" + 0.046*"maan"
2005	0.030*"suome" + 0.025*"maailm" + 0.022*"ilmastonmuutoks" + 0.021*"ihmis" + 0.021*"osa"
2006	0.012*"suome" + 0.008*"eu" + 0.007*"maailm" + 0.006*"ihmis" + 0.005*"usa"
2007	0.014*"ihmis" + 0.010*"suome" + 0.009*"maailm" + 0.006*"ilmasto" + 0.006*"ilmastonmuutoks"
2008	0.012*"suome" + 0.011*"kunt" + 0.008*"eu" + 0.007*"ilmasto" + 0.007*"ilmastonmuutoks"
2009	0.016*"suome" + 0.012*"maailm" + 0.009*"ihmis" + 0.006*"ilmastonmuutoks" + 0.006*"ilmasto"
2010	0.022*"ilmasto" + 0.015*"ihmis" + 0.014*"ilmastonmuutoks" + 0.012*"agw" + 0.011*"talv"
2011	0.020*"suome" + 0.016*"ihmis" + 0.011*"talv" + 0.011*"maailm" + 0.010*"lopu"
2012	0.038*"ilmasto" + 0.036*"suome" + 0.029*"talv" + 0.025*"näkyy" + 0.017*"kesä"
2013	0.034*"ilmasto" + 0.023*"näkyy" + 0.021*"suome" + 0.017*"ast" + 0.016*"ilmastonmuutoks"
2014	0.024*"maailm" + 0.019*"lopu" + 0.016*"automaatio" + 0.015*"the" + 0.013*"tot"
2015	0.032*"ihmis" + 0.030*"maailm" + 0.029*"suome" + 0.022*"osa" + 0.013*"ilmasto"
2016	0.019*"ilmasto" + 0.014*"ydinvoim" + 0.012*"maapalo" + 0.012*"ihmis" + 0.010*"ilmastonmuutoks"
2017	0.032*"ihmis" + 0.017*"aiheuttam" + 0.014*"suome" + 0.014*"ilmastonmuutoks" + 0.014*"maailm"
2018	0.023*"talouskasvu" + 0.014*"ikuis" + 0.012*"mere" + 0.010*"kasvihuonekaasu" + 0.009*"termostaat"
2019	0.019*"suome" + 0.017*"ilmasto" + 0.012*"ihmis" + 0.011*"ilmastonmuutoks" + 0.010*"maailm"
2020	0.016*"suome" + 0.014*"ilmasto" + 0.014*"ilmastonmuutoks" + 0.014*"ihmis" + 0.010*"maailm"

Fig. 9. Topic Modeling Analysis Results, Topic2 keywords

Analyzing the topic modeling results reveals fascinating shifts and trends in public discourse related to climate change. The discussion highlights these changes and notable topics:

*1) **2004-2006: Early Discussions and Emerging Awareness**:*

- **2004:** Focus on Finland's role in climate change and energy discussions. Emergence of the greenhouse effect as a topic indicates growing awareness.
- **2005:** Shift towards nuclear power and global perspectives, reflecting debates on energy solutions.
- **2006:** Presence of specific names suggests political involvement. EU and USA mentioned, indicating a broader geopolitical context.

*2) **2007-2010: Broadening Discussions and Global Events**:*

- **2007-2008:** Discussions become more general, focusing on climate and its changes.
- **2009:** Term 'maapalo' (wildfires) appears, indicating environmental events influencing discourse.
- **2010:** English language references indicate international influence. Continued relevance of environmental crises.

*3) **2011-2015: Deepening Understanding and Local Impact**:*

- **2011-2012:** Discussions about visible climate impacts in Finland.
- **2013-2015:** Focus on visible impacts, global warming, and economic aspects.

*4) **2016-2017: Political Influences and Global Leaders**:*

- **2016:** Discussions about neighboring countries' roles in climate issues.
- **2017:** Appearance of global political figures in climate change discourse with detection of keywords such as "trump".

*5) **2018-2020: Evolving Topics and Surprising Shifts**:*

- **2018:** Economic aspects and sustainability issues become prominent.
- **2019:** Discussion of political parties and the Green party, reflecting political dimensions.
- **2020:** Unexpected shift in topics, indicating a unique event.

*6) **Overall Observations**:* The evolution from specific energy discussions to broader climate change impacts and political influences reflects growing public awareness and the complexity of climate issues. The appearance of specific terms like wildfires and global political figures highlights how global events can influence local discussions.

G. Sentiment-Enhanced Temporal Topic Modeling

In this analysis, we utilized the AFINN sentiment analyzer to gauge sentiment scores for each statement and discussion. We further applied topic modeling to identify key topics associated with positive and negative sentiments in different years. "Fig. 10" to "Fig. 13" illustrate topic modeling keywords related to positive and negative threads.

Year	Topic 1 Keywords
2004	0.131*"ilmastonmuutoks" + 0.125*"suome" + 0.120*"energia" + 0.094*"käytö" + 0.063*"euroop"
2005	0.063*"suome" + 0.058*"ilmastonmuutoks" + 0.043*"osa" + 0.038*"ihmis" + 0.035*"ilmasto"
2006	0.012*"ilmasto" + 0.011*"ihmis" + 0.010*"alue" + 0.009*"käytö" + 0.008*"energia"
2007	0.020*"suome" + 0.010*"maailm" + 0.007*"suomi" + 0.007*"ihmis" + 0.006*"talv"
2008	0.014*"suome" + 0.012*"ihmis" + 0.011*"maailm" + 0.011*"ilmastonmuutoks" + 0.010*"ilmasto"
2009	0.022*"suome" + 0.011*"maailm" + 0.008*"ihmis" + 0.008*"osa" + 0.007*"pitä"
2010	0.027*"ilmasto" + 0.020*"ihmis" + 0.019*"suome" + 0.015*"ilmastonmuutoks" + 0.011*"maailm"
2011	0.019*"maailm" + 0.019*"ilmastonmuutoks" + 0.019*"osa" + 0.016*"ihmis" + 0.014*"lopu"
2012	0.080*"ilmasto" + 0.026*"talv" + 0.024*"suome" + 0.023*"the" + 0.023*"lämpenemis"
2013	0.032*"maailm" + 0.024*"lopu" + 0.023*"usko" + 0.018*"työpaik" + 0.015*"ydinas"
2014	0.036*"maailm" + 0.019*"ihmis" + 0.016*"ilmastonmuutoks" + 0.016*"automaatio" + 0.015*"lopullis"
2015	0.042*"ilmasto" + 0.041*"osa" + 0.039*"näkyy" + 0.022*"lämpenemin" + 0.019*"ilmastonmuutoks"
2016	0.043*"suome" + 0.033*"ilmasto" + 0.027*"ilmastonmuutoks" + 0.026*"maailm" + 0.025*"ihmis"
2017	0.030*"ilmasto" + 0.025*"ilmastonmuutoks" + 0.024*"trump" + 0.023*"ihmis" + 0.023*"aiheuttam"
2018	0.027*"talouskasvu" + 0.017*"ikuis" + 0.016*"kasvu" + 0.011*"ihmist" + 0.011*"ihmis"
2019	0.024*"ilmastonmuutoks" + 0.023*"ihmis" + 0.017*"ilmasto" + 0.010*"usko" + 0.010*"laps"
2020	0.034*"suome" + 0.021*"ilmasto" + 0.020*"talv" + 0.015*"ilmastonmuutoks" + 0.013*"yle"

Fig. 10. Positive Sentiment Analysis and Topic Modeling Results for topic1 keywords

Year	Topic 2 Keywords
2004	0.214*"uusiutuv" + 0.161*"suome" + 0.104*"energia" + 0.086*"kehity" + 0.075*"maan"
2005	0.116*"ydinvoim" + 0.078*"käytö" + 0.049*"raport" + 0.034*"climat" + 0.033*"ilmasto"
2006	0.020*"suome" + 0.016*"eu" + 0.011*"usa" + 0.010*"maailm" + 0.009*"euroop"
2007	0.017*"ihmis" + 0.017*"ilmasto" + 0.014*"ilmastonmuutoks" + 0.010*"maapalo" + 0.009*"the"
2008	0.017*"suome" + 0.013*"sähkö" + 0.011*"kunt" + 0.009*"energ" + 0.008*"energia"
2009	0.033*"ilmasto" + 0.020*"ihmis" + 0.016*"ilmastonmuutoks" + 0.011*"maapalo" + 0.009*"lämpötil"
2010	0.037*"the" + 0.022*"näkyy" + 0.021*"palst" + 0.017*"of" + 0.013*"ilmastonmuutoks"
2011	0.046*"ilmasto" + 0.026*"näkyy" + 0.025*"suome" + 0.022*"the" + 0.016*"ilmastonmuutoks"
2012	0.051*"näkyy" + 0.044*"ilmastonmuutoks" + 0.029*"ihmis" + 0.027*"ast" + 0.025*"suome"
2013	0.053*"ilmasto" + 0.036*"näkyy" + 0.028*"ilmastonmuutoks" + 0.025*"suome" + 0.024*"ast"
2014	0.048*"ilmasto" + 0.041*"näkyy" + 0.026*"talv" + 0.024*"suome" + 0.023*"ilmastonmuutoks"
2015	0.045*"ihmis" + 0.034*"maailm" + 0.031*"suome" + 0.026*"ilmastonmuutoks" + 0.015*"tutkij"
2016	0.021*"politiik" + 0.018*"eläm" + 0.017*"tieteellis" + 0.017*"miele" + 0.016*"päivä"
2017	0.039*"suome" + 0.039*"ihmis" + 0.035*"ydinvoim" + 0.032*"ilmastonmuutoks" + 0.024*"jumal"
2018	0.026*"ihmis" + 0.023*"suome" + 0.022*"ilmasto" + 0.021*"ilmastonmuutoks" + 0.013*"maailm"
2019	0.035*"suome" + 0.013*"maailm" + 0.012*"puolue" + 0.011*"ilmasto" + 0.009*"auto"
2020	0.026*"ihmis" + 0.021*"ilmastonmuutoks" + 0.020*"maailm" + 0.016*"ilmasto" + 0.012*"jumal"

Fig. 11. Positive Sentiment Analysis and Topic Modeling Results for topic2 keywords

Year	Topic 1 Keywords
2005	0.249*"suome" + 0.114*"tehd" + 0.100*"maan" + 0.092*"ilmastonmuutoks" + 0.087*"alue"
2006	0.019*"sähkö" + 0.017*"energia" + 0.016*"vaihd" + 0.015*"suome" + 0.012*"ilmasto"
2007	0.021*"ihmis" + 0.019*"maailm" + 0.012*"suome" + 0.011*"osa" + 0.008*"auto"
2008	0.018*"maailm" + 0.013*"suome" + 0.010*"kun" + 0.008*"valtio" + 0.007*"osa"
2009	0.031*"ihmis" + 0.017*"maailm" + 0.017*"ilmasto" + 0.013*"maapalo" + 0.009*"esim"
2010	0.029*"agw" + 0.026*"ihmis" + 0.023*"maapalo" + 0.019*"ilmastonmuutoks" + 0.014*"maailm"
2011	0.061*"the" + 0.049*"suome" + 0.044*"of" + 0.029*"and" + 0.029*"is"
2012	0.097*"ilmastonmuutoks" + 0.063*"suome" + 0.046*"ihmis" + 0.030*"usko" + 0.027*"maailm"
2013	0.054*"maailm" + 0.046*"lopu" + 0.031*"suome" + 0.030*"maapalo" + 0.023*"ihmis"
2014	0.032*"kiihtyy" + 0.028*"maailm" + 0.027*"lhc" + 0.027*"automaatio" + 0.025*"hawking"
2015	0.126*"suome" + 0.088*"maailm" + 0.046*"miele" + 0.043*"pohjois" + 0.039*"kansalais"
2016	0.057*"ilmasto" + 0.053*"ilmastonmuutoks" + 0.035*"valtio" + 0.029*"vuon" + 0.028*"miele"
2017	0.083*"suome" + 0.064*"sano" + 0.058*"maapalo" + 0.056*"ongelm" + 0.051*"ihmis"
2018	0.048*"suome" + 0.045*"ilmasto" + 0.025*"euroop" + 0.018*"alue" + 0.016*"keskustelu"
2019	0.023*"ihmis" + 0.022*"ilmasto" + 0.016*"maailm" + 0.015*"maapalo" + 0.015*"ilmastonmuutoks"
2020	0.026*"suome" + 0.022*"ihmis" + 0.021*"ilmasto" + 0.015*"ilmastonmuutoks" + 0.015*"tuho"

Fig. 12. Negative Sentiment Analysis and Topic Modeling Results for topic1 keywords

Year	Topic 2 Keywords
2005	0.135*"maailm" + 0.115*"toimin" + 0.101*"osa" + 0.097*"maapalo" + 0.080*"käyt"
2006	0.018*"suome" + 0.015*"ihmis" + 0.012*"nato" + 0.010*"eu" + 0.010*"niinistö"
2007	0.016*"ilmastonmuutoks" + 0.016*"ilmasto" + 0.012*"ihmis" + 0.011*"maapalo" + 0.008*"lämpötil"
2008	0.028*"ihmis" + 0.019*"ilmasto" + 0.016*"ilmastonmuutoks" + 0.015*"maapalo" + 0.012*"elām"
2009	0.017*"klo" + 0.017*"ilmastonmuutoks" + 0.016*"ihmis" + 0.016*"maailm" + 0.013*"suome"
2010	0.019*"the" + 0.017*"lopu" + 0.015*"suome" + 0.015*"talv" + 0.015*"ilmasto"
2011	0.036*"ihmis" + 0.026*"ilmastonmuutoks" + 0.025*"ilmasto" + 0.023*"ihmiskun" + 0.022*"maailm"
2012	0.041*"ihmis" + 0.039*"ilmasto" + 0.039*"vuon" + 0.037*"maapalo" + 0.036*"vaikut"
2013	0.077*"ilmastonmuutoks" + 0.056*"ilmasto" + 0.035*"ihmis" + 0.029*"denialist" + 0.027*"näkyy"
2014	0.079*"ihmis" + 0.068*"ilmasto" + 0.060*"maailm" + 0.047*"suome" + 0.038*"ilmastonmuutoks"
2015	0.137*"ihmis" + 0.092*"ilmasto" + 0.047*"venäj" + 0.034*"saat" + 0.032*"osa"
2016	0.047*"suome" + 0.041*"ydinvoim" + 0.040*"ihmis" + 0.039*"maailm" + 0.038*"venäj"
2017	0.070*"ilmasto" + 0.064*"ihmis" + 0.055*"uhk" + 0.046*"trump" + 0.041*"aiheuttam"
2018	0.028*"ihmis" + 0.026*"maapalo" + 0.020*"maailm" + 0.017*"mere" + 0.017*"ilmastonmuutoks"
2019	0.023*"suome" + 0.021*"ihmis" + 0.015*"sipil" + 0.013*"maailm" + 0.012*"maaseudu"
2020	0.041*"kenkäpar" + 0.026*"keng" + 0.021*"kuoma" + 0.016*"minu" + 0.015*"valmistaj"

Fig. 13. Negative Sentiment Analysis and Topic Modeling Results for topic2 keywords

*1) **Themes and Focus**:*

- *Positive Sentiments:* Discussions under positive sentiments predominantly revolve around solutions, advancements in renewable energy, and proactive measures against climate change. Terms like 'uusiutuv' (renewable) and 'energia' (energy) in 2004, and 'näkyy' (visible) in later years, indicate a hopeful perspective.
- *Negative Sentiments:* In contrast, negative sentiments often highlight skepticism, political controversies, and challenges in tackling climate change. Terms like 'denialist' in 2013 and 'trump' in 2017 underscore the politicization of climate change discussions.

*2) **Evolution of Discussion Over Time**:*

- *Positive Sentiments:* There is a shift from general awareness to more specific topics, such as the impact of automation in 2014 and local issues in 2020, reflecting a deepening understanding of climate change.
- *Negative Sentiments:* Negative discussions evolved from focusing on Finland's role to encompassing global skepticism and political factors, with terms like 'venäj' (Russia) in 2016 and 'trump' in 2017.

*3) **Specific Mentions and Trends**:*

- In positive sentiments, the consistent mention of 'ilmastonmuutoks' (climate change) and 'suome' (Finland) across years highlights a sustained focus. The appearance of 'talv' (winter) and 'lämpenemis' (warming) suggests acknowledgment of climate effects.
- In negative sentiments, the shift from localized concerns to global political issues reflects the expanding scope of climate change discourse.

H. Topic Modeling Using Non-Negative Matrix Factorization (NMF)

NMF method is used for the topic modeling of all threads to compare the results and provide more comprehensive results. The topic keywords generated by this algorithm were shown in "Fig. 14" and "Fig. 15".

Year	Topic 1 Keywords
2004	suome energia hiilidioksid polto ilmastonmuutoks
2005	ydinvoim käytö iaea raport the
2006	suome hallituks eu ihmis maailm
2007	ihmis ilmasto ilmastonmuutoks maapalo maailm
2008	ihmis ilmasto suome maailm ilmastonmuutoks
2009	ihmis ilmastonmuutoks suome maailm osa
2010	ilmasto ihmis ilmastonmuutoks agw talv
2011	ilmasto suome ihmis ilmastonmuutoks talv
2012	ilmasto suome ilmastonmuutoks ihmis talv
2013	ilmasto näkyy suome ast ilmastonmuutoks
2014	näkyy ilmasto suome talv lämpenemin
2015	ihmis maailm ilmastonmuutoks suome ilmasto
2016	ilmasto suome ilmastonmuutoks ihmis näkyy
2017	suome ilmasto ihmis ilmastonmuutoks näkyy
2018	suome ilmasto ilmastonmuutoks ihmis maailm
2019	suome ilmasto ihmis maailm ilmastonmuutoks
2020	ilmasto ilmastonmuutoks suome ihmis maailm

Fig. 14. Annual Topics Extracted Through NMF Topic Modeling for topic1 keywords

Year	Topic 2 Keywords
2004	selonteo uhk suome sotilaallis main
2005	huom turvallisuus katso viisas niinistö
2006	vaihd sähkö energia tuot valits
2007	vastu kansalaist ylittäv suomi suome
2008	vastu kansalaist hallitusohjelm hallitus suomi
2009	ilmasto ast maapalo muutos lämpenemin
2010	näkyy palst keskustel record vaikutuks
2011	näkyy the in of and
2012	näkyy lämmin ennätysl no jäätikö
2013	lopu maailm työpaik lähestyy ydinas
2014	maailm automaatio lopu lopullis työpaik
2015	näkyy lämpenemin sate ilmasto tulo
2016	natsism uusnats huuha politiik tieteellis
2017	trump president ihmist yhdysval usa
2018	talouskasvu ikuis termostaat kasvihuonekaasu mere
2019	nostamin hillitsemis maitotuot vero ilmastonmuutoks
2020	kenkäpar keng kuoma jalkin valmistaj

Fig. 15. Annual Topics Extracted Through NMF Topic Modeling for topic2 keywords

*1) **NMF Method**:*

- **General Trends:**
  - The NMF method highlights climate change and environmental issues, focusing more on societal and political aspects, like 'turvallisuus' (safety) in 2005 and 'hallituks' (government) in 2006.
  - There's a clear presence of discussion around human impact and societal responsibility'ihmis', 'vastu').
- **Interesting Observations:**
  - The NMF results show a focus on current events and political figures, such as "trump" in 2017.
  - Terms like 'natsism' and 'huuha' (nonsense) in 2016 indicate polarized or controversial discussions.

*2) **Differences Between LDA and NMF Outputs**:*

- **Topic Coherence and Specificity:**
  - LDA produces broader topics, while NMF results in more specific and coherent topics.
  - LDA captures overarching themes, NMF pinpoints specific issues or events.
- **Sensitivity to Language and Context:**
  - LDA is less sensitive to specific language, while NMF is more contextually aware.
- **Temporal Dynamics:**
  - LDA shows gradual evolution of topics, NMF is responsive to specific yearly changes.

*3) **Why These Differences Occur**:*

- LDA is a probabilistic model suitable for identifying overarching themes. LDA is suitable for understanding broad themes.
- NMF, based on linear algebra, detects specific patterns and nuances, making it sensitive to specific word associations and context. NMF is appropriate for specific, nuanced discussions.

Both methods have strengths and can be used in conjunction for a comprehensive understanding.

V. Conclusion

In conclusion, our analysis of the Suomi24 corpus from 2001 to 2020 has provided significant insights into the linguistic patterns and thematic evolution of discussions around climate change in Finnish online forums. Through the application of Zipf's and Heaps' laws, we have quantitatively tracked the growth and diversification of vocabulary, reflecting the dynamic nature of public discourse. The lexical distance analysis has revealed the contextual framework of climate change discussions, highlighting the most frequently co-occurring terms. Furthermore, sentiment analysis using the AFINN tool has offered a perspective on the emotional dimensions of these discussions. The use of LDA and NMF topic modeling has been pivotal in identifying and tracking the main themes over the years, illustrating the shifting focus and concerns in public discourse. This study not only contributes to our understanding of language use in online forums but also provides a valuable lens through which to view public perception and discourse on climate change over two decades.

References

[1] J. Tana, E. Eirola, and K. Eriksson-Backa, "Exploring temporal variations of depression-related health information behaviour in a discussion forum: The case of suomi24," 2020.

[2] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic bulletin & review*, vol. 21, pp. 1112–1130, 2014.

[3] D. C. van Leijenhorst and T. P. Van der Weide, "A formal derivation of heaps' law," *Information Sciences*, vol. 170, no. 2-4, pp. 263–272, 2005.

[4] E. Haapamäki, J. Mikkola, M. Hirsimäki, and M. Oussalah, "Mining security discussions in suomi24," in *2019 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 2019, pp. 101–108.

[5] M. S. Jahan, M. Oussalah, and N. Arhab, "Finnish hate-speech detection on social media using cnn and finbert," in *Language Resources and Evaluation Conference, LREC 2022, 20-25 June 2022, Palais du Pharo, Marseille, France: conference proceedings*. European Language Resources Association, 2022.

[6] M. Ibrahim, M. Eteläperä, S. Turkmen, M. Maged, M. Oussalah, and J. Miettunen, "Mining health discussions on suomi24," in *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2019, pp. 1580–1585.

[7] J. Tana, A. Shcherbakov, and L. Espinosa-Leal, "Sentiment analysis of depression related discussions in the suomi24 discussion forum," *Informaatiotutkimus*, vol. 41, no. 2–3, pp. 157–162, 2022.

[8] J. Tana, E. Eirola, and K. Eriksson-Backa, "The aspect of time in online health information behaviour: a longitudinal extensive analysis of the suomi24 discussion forum," *Informaatiotutkimus*, vol. 38, no. 2, pp. 7–31, 2019.

[9] I. Moamen, M. Oussalah, D. Ackers, E. Gilman, and J. Miettunen, "Psychosocial factor identification in cancer patient community," 2020.

[10] Aller Media Ltd., "The Suomi24 Sentences Corpus 2001-2017, Korp version 1.2," 2019-01-01. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2020021803

[11] City Digital Group, "The Suomi24 Corpus 2018-2020, VRT version," 2021-01-01. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2021101523

[12] A. Cichocki and A.-H. Phan, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 92, no. 3, pp. 708–721, 2009.

[13] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[14] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo, "Multilingual is not enough: Bert for finnish," *arXiv preprint arXiv:1912.07076*, 2019.

APPENDIX

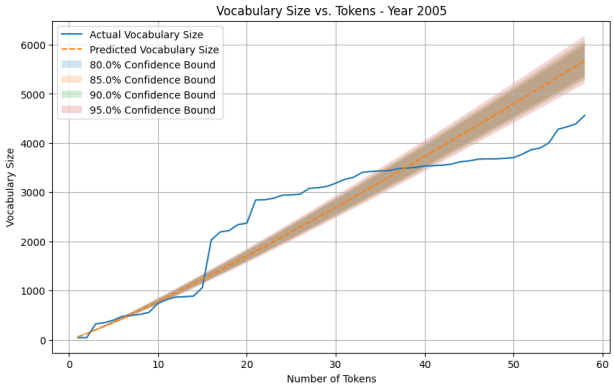

Fig. 16. Vocabulary Growth and Heaps' Law Fit in the Year 2005

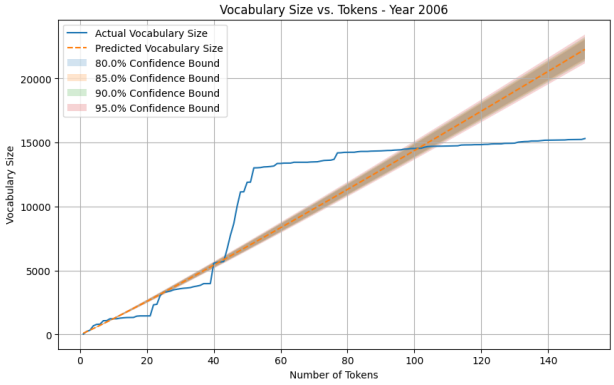

Fig. 17. Vocabulary Growth and Heaps' Law Fit in the Year 2006

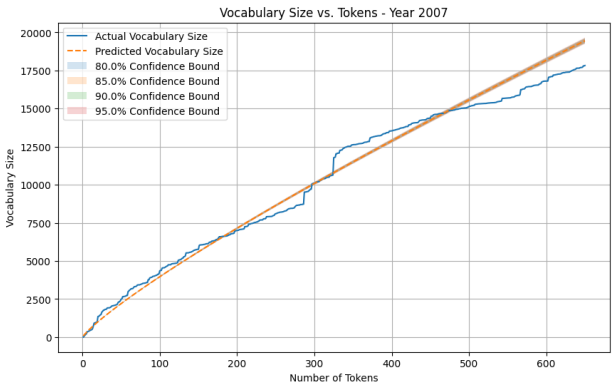

Fig. 18. Vocabulary Growth and Heaps' Law Fit in the Year 2007

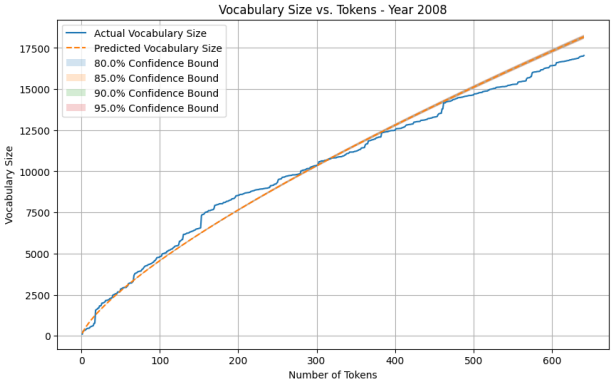

Fig. 19. Vocabulary Growth and Heaps' Law Fit in the Year 2008

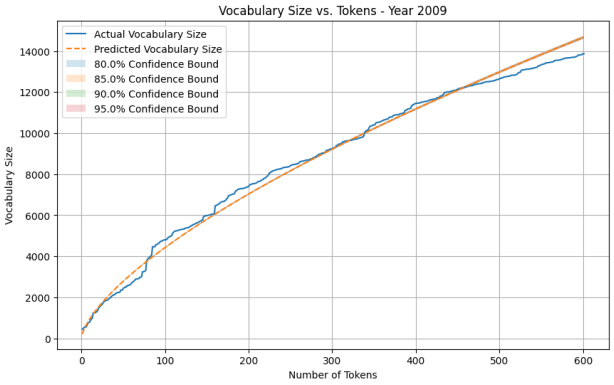

Fig. 20. Vocabulary Growth and Heaps' Law Fit in the Year 2009

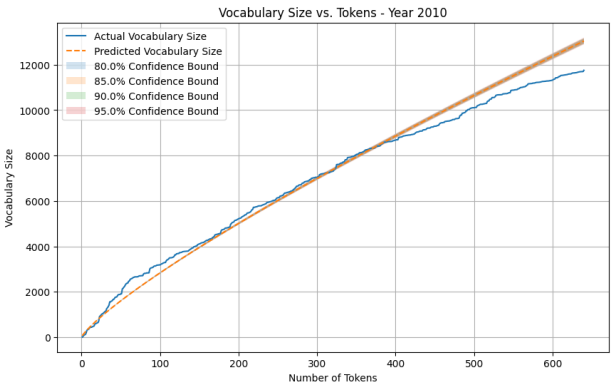

Fig. 21. Vocabulary Growth and Heaps' Law Fit in the Year 2010

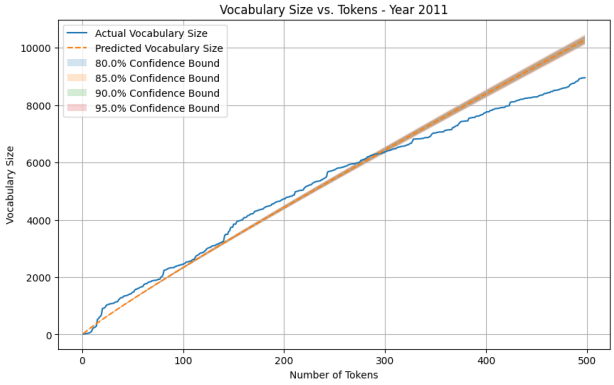

Fig. 22. Vocabulary Growth and Heaps' Law Fit in the Year 2011

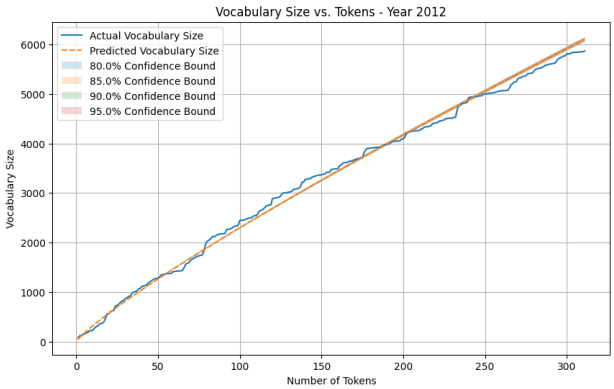

Fig. 23. Vocabulary Growth and Heaps' Law Fit in the Year 2012

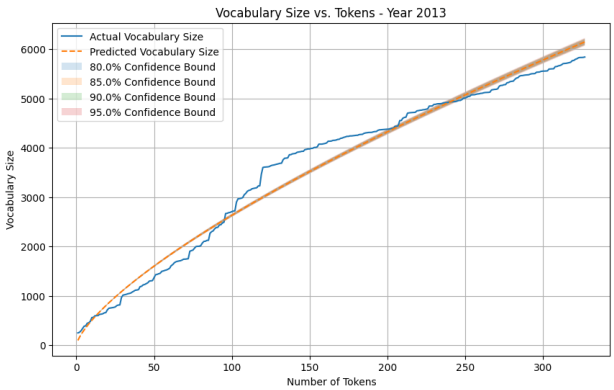

Fig. 24. Vocabulary Growth and Heaps' Law Fit in the Year 2013

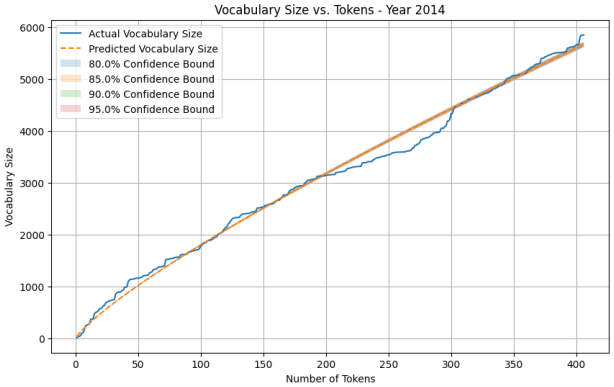

Fig. 25. Vocabulary Growth and Heaps' Law Fit in the Year 2014

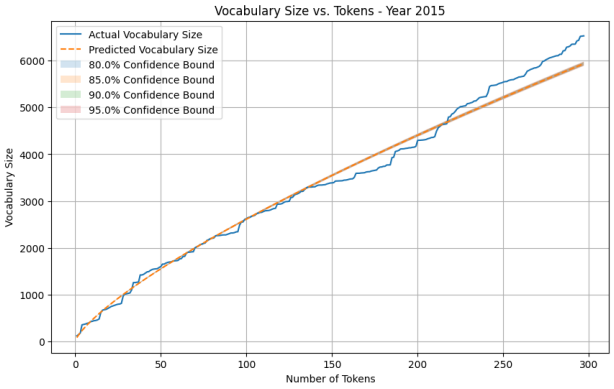

Fig. 26. Vocabulary Growth and Heaps' Law Fit in the Year 2015

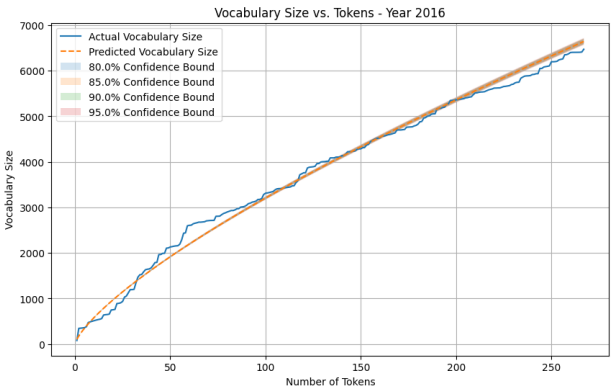

Fig. 27. Vocabulary Growth and Heaps' Law Fit in the Year 2016

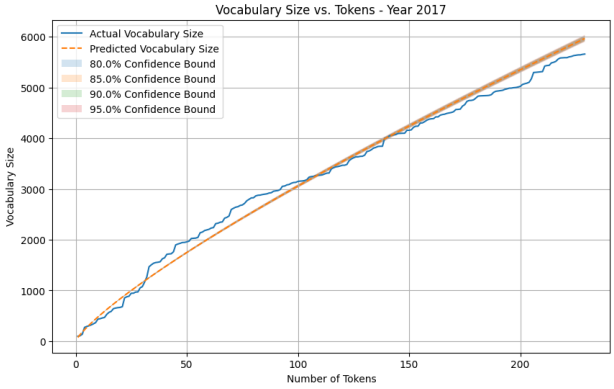

Fig. 28. Vocabulary Growth and Heaps' Law Fit in the Year 2017

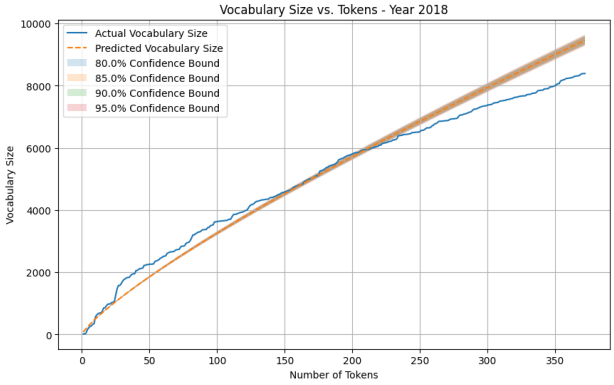

Fig. 29. Vocabulary Growth and Heaps' Law Fit in the Year 2018

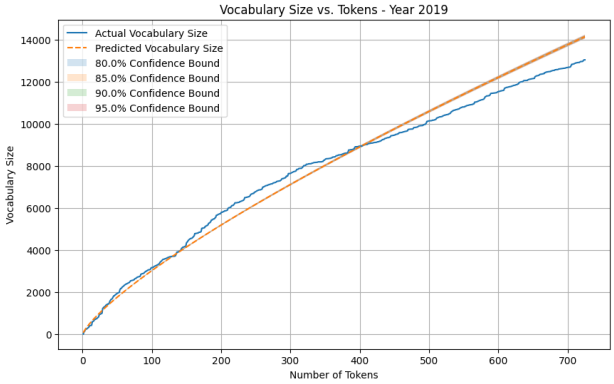

Fig. 30. Vocabulary Growth and Heaps' Law Fit in the Year 2019