KØBENHAVNS UNIVERSITET

DEPARTMENTOFECONOMICS

Exploring the Semantics of Fiscal Sustainability in Greenland

Contribution:

Examination number 42: 1, 2, 3, 4,1, 4.3, 4.4.2, 4.4.5, 5, 5.1

Examination number 13: 1, 2, 3.1, 4,2, 4.4, 4.4.3, 4.4.6, 6

Examination number 118: 1, 2, 3.2, 4.2.1, 4.4.1. 4.4.4, 4.5, 7

Keystrokes incl. spaces: 34,218

Submission date: 24/08/21

TABLE OF CONTENTS

1. INTRODUCTION

In later years the emergence of computational ability within social science and the expansion of the world wide web has propelled the amount of available digital data sources and provided novel opportunities to explore, analyze and understand human language via computational activities such as natural language processing. Recently there has been a larger focus on the semantic and syntactic meaning of words generated by word embeddings via representation of text in a dimensional vector space (Mikolov et al., 2013). These vectors of words can be utilized to calculate the meaning and context of each word based upon the similarities and differences of the vector distributions. Often, these digital methods have been utilized to explore the emotional meaning behind words, as for example gender biases (Bolukbasi et al. 2016; Vágsheyg & Wilhelmsen 2018), the historic semantic development of words (Hamilton et al., 2016) and real-life personalities via social media content (Arnoux et al. 2017).

In this paper, we are utilizing word embedding for a different type of digital data source and goal. Instead of focusing on a certain bias or associated emotions, we are focusing on professional texts within the budgetary aspects of Government finances. Namely, we want to explore the semantics associated with five preconcepted themes across two government outlets in Greenland, the Economic Council, and the Department of Finance, with varying degrees of dependence and association with Greenlandic self-governance. Thus, we seek to answer the following research question:

What characterizes the semantics associated with fiscal sustainability in economic reports?

We consider several findings. We manage to present a novel way of exploring professional economic reports via visualization of word embeddings. Next, we explore a semantic difference across two different state-associated economic actors and seek explanation via a provided preconception regarding the actors. The paper is organized as follows: Section 2 presents an overview of the existing literature. In section 3, the paper gives an overview of Word embeddings, including the Word2Vec model. In section 4, the data is described and how this paper works with data wrangling and the empirical strategy. Section 5 presents the empirical results. Finally, section 6 and 7 summarizes the main findings and concludes.

2. LITERATURE REVIEW

This section provides a brief overview of contemporary scientific use of word embeddings. Word embeddings have been used for a wide variety of scientific purposes. Our literature has mainly been found using Google Scholar and have been sorted by citations in order to achieve relevance. Examples of usage are exploration of biases in news and job applications (Bolukbasi et al., 2016; Vágsheyg & Wilhelmsen, 2016), changing semantics of words throughout history (Hamilton et al., 2016) and classifying real-life personality via twitter-usage (Arnoux et al., 2017).

Bolukbasi et al. (2016) and Vágsheyg & Wilhelmsen (2016) find word embeddings are well suited for discovering latent biases in news and job applications. Both find that word embeddings contain biases that reflect gender stereotypes in broader society. Namely, certain occupations are implicitly associated with the female and male gender in both job advertisements on the danish job board Jobindex and Google News articles. This knowledge is useful for developing a debiazing algorithm while preserving the utility of the embedding, to provide a fairer set of inputs for machine learning tasks.

Hamilton et al. (2016) finds promise utilizing word embeddings as a diachronic tool, helpful to understand how words change their meaning over time. They develop a robust methodology for measuring semantic change by evaluating word embeddings against known historical changes.

Arnoux et al. (2017) demonstrates that the use of word embedding can predict users' Big-5 personality traits from their social media text in a real-life context. They improve on former personality modelling by significantly reducing the text size requirement and therefore more applicable to real life social media context.

The state of the art regarding research into the utility of word embeddings have therefore utilized data sources from news outlets, job postings, historical datasets of books and social media among others. It is thus our proposal that this field of research can benefit from our inquiry into another type of digital source, namely professional texts regarding the economic development of a country and the topics of salience hereof.

3. THEORETICAL FRAMEWORK

This section presents the theoretical framework of the paper. Word embeddings is gaining an increasing popularity, currently the models are based solely on linear contexts. In this work, we generalize the Word2Vec method introduced by Mikolov et al. (2013) to include arbitrary contexts, this application is a technique to learn word embeddings using shallow neural networks.

3.1 WORD EMBEDDINGS

Word embeddings is an effective tool for the dense representations of individual words in a text. This tool enables us to capture the context of a word when analyzing it as a fragment from a text and draw upon the semantic and syntactic similarity and find the relation between other words. Word embeddings is a class of techniques where individual words are represented as real-valued vectors in a predefined vector space. Every single word is mapped to a vector and the vector values are taught in a way that resembles a neural network. The embeddings are well-known for being the distributed language representations, map from sentences to a discrete path through a latent vector space. The vector space, V, consists of an embedding (a location) for every vocabulary word in K, where K is the dimension of the latent representation space. The embeddings are selected to improve, approximately, an objective function defined on the original text, such as the likelihood of word occurrences (Gentzkow et al., 2019).

The main approach is the idea of using a dense distributed representation for each word. In this paper, we perform experiments with five pre-selected themes by using Word2Vec, this tool is a statistical method for efficiently learning a standalone word embedding from a text corpus.

3.2 WORD2VEC

Word2vec provides a unique perspective to the text mining literature and by converting words and phrases into a vector representation, as a result of this tool, each vector representation for a given word is trained to be highly probable given the vector representation of the surrounding context. Assume $w_{sj} \in \{1 \ldots p\}$ denotes the identity of the j^{th} word in sentence s. The embedding algorithms can be solved by the following:

$$\arg\max_{\mathbf{V},\mathbf{U}} \sum_{s,t} \sum_{j \neq t,\ j=t-b}^{t+b} \mathbf{v}'_{w_{sj}} \mathbf{u}_{w_{st}} - A(\mathbf{V},\mathbf{U}) \tag{1}$$

For equation (1), V and U are $K \times p$ matrices with columns (e.g., $v_{w_{sj}}$) that denote the embedding of each word in a vector space. Word2Vec is implied by the multinomial likelihood, where $A(\cdot, \cdot)$ is a normalizing function and there are two embedding spaces estimated, V and U. Most often, these will be near mirror images of each other, and only one space will be reported, but in general, these two spaces and the angle between words can be useful (Gentzkow et al., 2019). Many researchers connect these vector-space language models with the sorts of document attributes that are of interest in social science. Mikolov (2014) estimate latent document scores in a vector space, while Taddy (2015) develops an inversion rule based on Word2Vec for document classification. Based on the assumption that Word2vec adds extra semantic features that help in text classification. This paper pursues to demonstrate the effectiveness of visualizing Word2Vec generated from the open-source Python library Gensim.

4. DATA

4.1 ETHICAL CONSIDERATIONS IN DATA COLLECTION

Data collection is a central part of this paper, and the collecting of data is often done under the assumption that information provided is confidential and the findings will be anonymous. The Government of Greenland does not contain any sensitive personal data thus we do not have to consider GDPR. There are no ethical concerns given our data is obtained from an open API and we are using publicly available data (Naalakkersuisut 2013). When applying the Government of Greenland's API, we get access to all published data. We can apply the open data directly in various applications. A developer account is not a requirement, as www.naalakkersuisut.gl/en gives us an overview of the open data that the Government of Greenland provides to the public. The data can be downloaded in table format, graphics, maps, several data formats, and bulk downloads by an API entry point (Salganik, 2017).

4.2 DATA SOURCES

In order to explore the semantics associated with Greenlandic financial communication, we utilize data from the annual economic reports of the Economic Council Greenland and the Financial Department of Greenland. The reports from these different state actors are suitable for a multiple of reasons: (i) the government of Greenland has made the data publicly available via their homepage, (ii) it allows us to explore potential semantic differences in economic saliences between the department of a ministry and an independent council organ, and because (iii) exploring the word embeddings are relatively well suited as both reports are using the same economic terminology. This section provides a description of both organs, their reports and the preconcerted difference in output.

4.2.1 DESCRIPTION OF ACTORS

The Economic Council Greenland (EC) is an independent actor closely associated with the Ministry for Finance and Domestic Affairs of Greenland. The councils' duties consist of providing the ministry with regular appraisals via reports of the economic state of Greenland and estimate the sustainability of the pursued fiscal policies (Economic Council, 2021). Regarding the EC's reports, the first chapter usually focuses on the apparent economic trends and state of the economy for the past year, while the second chapter usually highlights the sustainability of the government's fiscal policy (Economic Council, 2021).

The other actor is the Department of Finance Greenland (DF). It serves the role of a secretariat to the ministry and has the responsibility of making sure the fiscal policy can be carried out via the annual Finance Act Bill. It has the responsibility of publishing "The Political Economic Report", which outlines and analyzes economic developments within the Greenlandic economy and activities in key areas such as expenditure and revenue and describes the key areas guiding the government's economic policy, including preparation for the coming years Finance Act Bill, therefore highlighting particular areas of concern and trends for the Government of Greenland (Naalakkersuisut Beretning, 2021).

4.3 Understanding the differential focus of the Economic Council and the Finance Department

To understand both the EC and the FD, we can compare their structural ties to the Government of Greenland.

The FD and the EC have several similarities, which is beneficial to overview before diving into the differences. Both organs serve a role as part of the Greenlandic self-governance. Both are deployed by the Ministry of Finance. While the department is serving directly under the ministry, the Economic Council also consists of a chairmanship appointed by the government, and several members from the government administration (Departementet for Finanser, 2021)

The yearly Finance Act bill process starts out with FD's preparation of the Political Economic Report. The Finance Department is therefore a direct hierarchical actor under the principal ministry and produces the Political Economic Report in order to describe the contemporary economic status and the government's expected prioritizations the following financial year. On the accounts of these structural dependencies, the FD's reports have a more short-sided and policy-oriented focus (Finansloven Naalakkersuisut, 2021).

The EC on the other hand, with its status as an independent advisory, represents a break with the direct hierarchy of the parliamentary chain of governance, and therefore achieves a higher level of structural independence (Blom-Hansen et al., 2014). This allows the council to focus more on the long-term sustainability of the fiscal policy and consider structural issues such as demographic vulnerability. A term often used throughout the EC's reports is *fiscal sustainability* ('Finanspolitisk holdbarhed'). Long term economic development is a highly speculative outlet, but the independent state of the council allows it to work with long-term assumptions such as demographic trends in order to forecast the economic development, which again serves to separate it from the more goal-bound, short-sighted outlet of the Finance Department.

Table 1 - **Overall themes and keywords**

Finance	Assessment	Time	Business	Trends
finanspolitisk holdbarhed finanspolitik udgiftspolitik	fremgang tilbagegang sunde uhensigtsmæssig ansvar efterslæb	langsigtet kortsigtet konjunktur struktur	råstofefterforskning fiskeri turisme	uddannelsesindsats erhvervsstruktur bosætningsmønster infrastruktur sundhedsområde

Based upon this aforementioned knowledge and the theoretical preconceptions of the above actors, we chose to operationalize our vectoring around five overall themes. The theme "Finance" is of the greatest importance. It allows us to explore the associations with fiscal sustainability and allows us to draw comparisons between the EC and FD. "Assessment" allows us to explore the potential negative or positive assessment of the aforementioned politics. "Time" allows us to explore the long- or short-term focus, hereof focus upon structural problems. "Business" allows us to explore what some of the most salient sectors of the Greenlandic economy are associated with. "Trends" should allow us to explore some of the most important developmental trends of the Greenlandic economy.

To maintain transparency, we are disclosing how these were decided upon. The themes have not only come about via our knowledge of the EC and FD, but also via an iterative process whereof several words have been tested and tried. However, these themes were not changed throughout our coming visual and comparative analysis.

4.4 Data Wrangling

In the following section, we will review our data wrangling that was performed in a Jupyter Notebook (cf. the associated Jupyter Notebook). Our goal is to turn annual financial reports in readable pdf format, available online, into a Python object that can be loaded into Gensim's Word2Vec model. More specifically, this object consists of a list of documents, which itself consists of lists of words (Rehurek 2021). A nested list of words and documents in other words. However, it is important to note that we are only interested in words that can act as features for word embeddings. This means that we not only have to retrieve and prepare data in a proper manner, but also modify the data in a way that we find relevant for an analysis.

4.4.1 MAPPING

We are interested in the annual reports of the EC and the Annual Political-Economic Report of the DF. All reports are readable pdf files available on the Greenlandic Government's website. Hence, we need to locate the URLs associated with each report to retrieve the data. Fortunately, the URLs are fixed for the EC and change only with the years of publication. For this reason, we can automate some of the mapping by utilizing a lambda function and a f-string. In contrast, the URLs of the Ministry of Finance appear arbitrary and are collected manually. We store all our URLs as elements in a list.

4.4.2 DOWNLOAD

We use the Request package to download the reports. In practice, we use a function or code snippet from Stack Overflow that takes a URL and a filename as input and downloads a PDF as output (Stackoverflow, 2021). To minimize coding, we zip our list of URLs along a list of appropriate filenames, for which each element is looped through the downloading function. The result is a set of downloaded PDF files in our working directory.

4.4.3 PARSE

Before we can extract the text data from the reports in our working directory, we need to load the PDFs in Python. We use the PyPDF2 package to both load the pdfs and extract the text data. It is worth mentioning that there are several packages that process pdfs in Python (LibHunt, 2021), but we have chosen PyPDF2 because it is well documented and fits our needs (Phaseit Inc and Fenniak, 2016). First, we load each pdf with PyPDF2's pdf reader and store each loaded pdf as an element in a list. Subsequently, we create a function with PyPDF2's page extractor that fills an empty string with text for each page in a pdf. In this way, we can apply the text extractor function on the list of loaded pdfs and store the strings as elements in a new list.

4.4.4 COMPLICATIONS OF RETRIEVING DATA

Ideally, we could have retrieved 22 annual reports from 2010-2020, but due to various complications we have only retrieved 15 reports in total, including 6 reports from the Economic Council and 9 reports from the Department of Finance. Regarding the Economic Council, we exclude reports from 2010-2011, as these reports are written in both Greenlandic and Danish. In addition, Py2PDF cannot load the reports from 2012-2015 for unknown reasons. This is probably due to a bug in the pdf reader, or something related to the file formats. Consequently, we are not picking up the hype around the mining industry, which was at its peak around 2012-2015 (Råstofeventyret, 2013). This may become important when comparing word embeddings across the actors, however, it is difficult to assess what impact this temporal aspect will have on the word embeddings itself. Regarding the Department of Finance, we find only a financial statement from 2018, which we discard, and a dead link for 2013. This leaves us with EC reports for the period 2016-2020 and DF reports from 2010-2020 except 2013 and 2018.

4.4.5 PREPROCESSING

Our preprocessing consists of segmentation, standardization, and removal of stop words. Basically, these processes help us transform our raw text data into a meaningful vocabulary that we can analyze for later. In practice, we create a function that lemmatizes each word in a list of strings that are split into lists of words. During our segmentation, i.e. splitting the strings into lists of words, we also set the words to lowercase and sort out punctuation and digits. Furthermore, our lemmatization only takes place for words that do not appear in a list of stop words. Finally, we also form relevant bigrams from our vocabulary. This is done manually by finding consecutive words and joining them together if they meet some specified criteria. In practice, we transform our list of words into bigrams using the nltk package's ngram function. If the criteria are not met for the individual bigram, we remove the secondary part of the bigram and delete the element in front of all the retained bigrams, since these elements are duplicates. We delete the duplicates in a descending order to preserve the index number of the next element for removal. As an example, we form the bigram *finanspolitisk holdbarhed* (fiscal sustainability) since this is an economic concept that is different from the words "fiscal policy" and "sustainability" per se (De Økonomiske Råd, 2012). Our preprocessing is summarized as follows.

Table 2 - **Data summary**

Actor	No. of Reports	No. of Raw Words	No. of Words	No. of Unique Words
Greenland Economic Council	6	93,286	41,102	11,224
Department of Finance	9	143,788	65,517	17,080
Total	15	237,074	106,619	28,304

Overall, our word count drops from 237,074 to 108,876 words, including 28,666 unique words.

4.4.6 Complications of modifying data

Our preprocessing does a noticeable job on our text data but leaves no guarantee of working as intended. This is because our lemmatizer does not work with 100 percent accuracy and the fact that our list of stop words is not complete. We use a pretrained lemmatizer from GitHub called "Lemmy". Lemmy is trained on data from both Danish and Swedish language advisory committees (Lemmy, 2019). Due to limited Danish resources, we have chosen to use Lemmy because it works as a standalone and easy to use lemmatizer. Although Lemmy is not perfect, we believe that this kind of standardization is better than nothing. Our list of stop words is also found on GitHub (Stopwords, 2021). An obvious alternative is the nltk package's list of Danish stop words, but we have chosen to use the list from GitHub, since this list is longer than the nltk version (262 versus 94 words). As a final step, we inspect our preprocessed list of words for additional stop words that may have slipped through the processing. This includes frequent names of board members and meta text such as "chapter" and "summary".

4.5 Empirical strategy

In short, our objective is to visualize word embeddings from annual economic reports by the Economic Council and Ministry of Finance. We use the Word2Vec Model from the Gensim package to make the word embeddings. The visualization itself is inspired by Gensim's own documentation on the same matter (Rehurek, 2021). This involves dimension reduction using t-distributed Stochastic Neighbor Embedding (t-SNE) by Scikit-learn. This is obviously a complex area, but we choose to follow Gensim's instructions. The basic idea is to transform each word in our vocabulary into a set of numeric vectors in k dimensions. Subsequently, our vectors can be

visualized in an ordinary coordinate system by transforming them into vectors in 2 dimensions. In this way, the axes are not interpretable in the same way as with a regression plot. It must be stressed that our visualization depends heavily on the methodological choices mentioned above. We set the vector size to 200 dimensions in a continuous-bag-of-words model and sort out words with frequency less than 5. We are also aware that a "window" parameter is set to 5 by default. This specifies the maximum distance between the target word and the surrounding words. In terms of reproducing our results, it should also be noted that our implementation of t-SNE uses the seed number 80720 (the wedding day of a group member). Hence, a different specification leads to a different visualization.

5. EMPIRICAL ANALYSIS

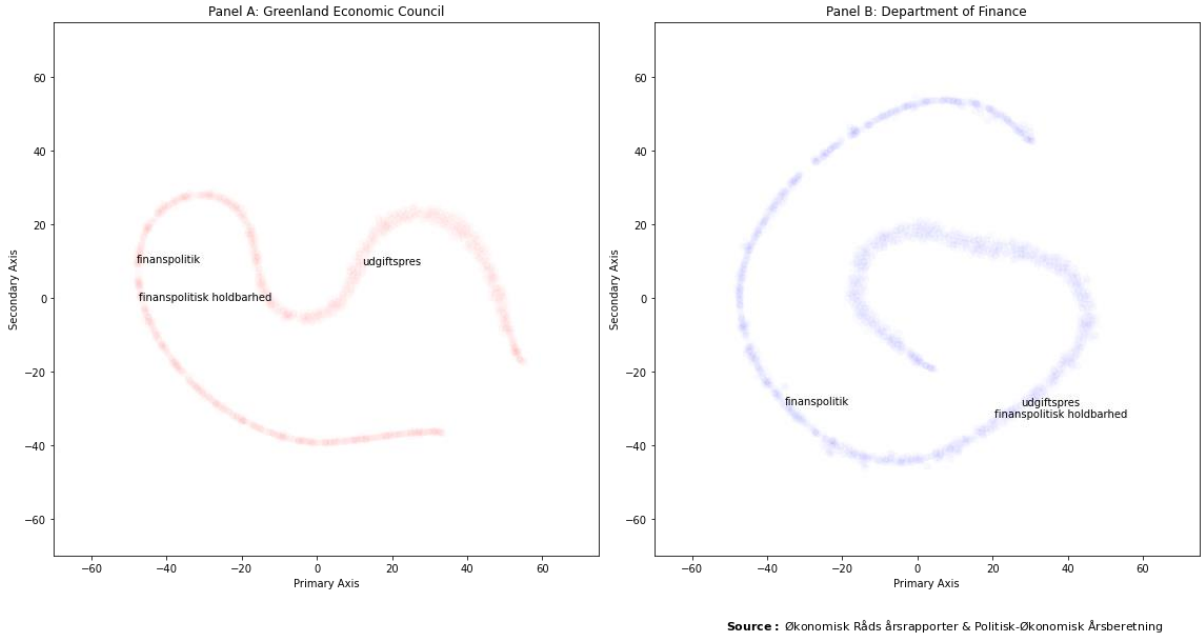
In the following analysis, we present our visualization of the semantic associations regarding Finance and our themes. Initially we will present a short guideline for reading our graphs. Next, we will analyze the meaningful distances between the keywords in each theme, while providing a brief comparison between the GEC and the DF.

5.1 VISUAL ANALYSIS OF THE SEMANTIC FEATURES

We plot all our word embeddings for the EC and DF separately, confront panel A and Panel B in figure 0. This means that the snake-like expressions in each panel are vector representations of the entire vocabulary collected from the EC and the DF. Once again, we must stress that the visual expression is a consequence of our methodological choices. If we had chosen to work with a skip gram model or PCA, our visual expression would have been a regular point cloud. Thus, each point represents a single word from our vocabularies. We have muted the colors by making the points transparent. In this way, annotated words appear clearly when we interpret their relative positions, since the distances represent semantic similarity.

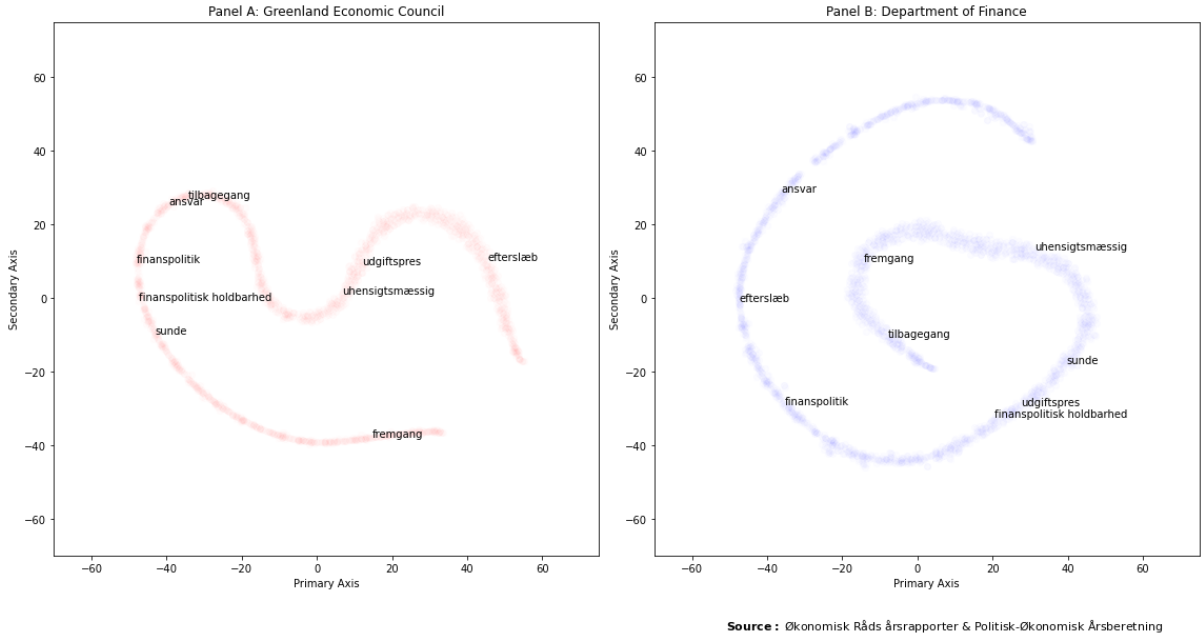
As aforementioned in section 4 about data sources, we will conduct our explorative analysis based upon a preconception about our actors. The main theme, in which we will have our semantic focus, is therefore "finanspolitik", "finanspolitisk holdbarhed" and "udgiftspolitik" (See section 4 and table 1 for an overview of total keywords).

Figure 0 - **Finance and sustainability**



The two graphs highlight the semantic associations regarding our main theme *finance*. Initially, one can observe the distances between "finanspolitik" and "finanspolitisk holdbarhed". This serves to explore our initial preconception regarding the semantics surrounding sustainability. For the GEC, the distance between the two vectors is noticeably shorter compared to the DF. This could potentially illustrate that sustainability shares more semantic similarity with fiscal policy for the GEC, while the DF shares more semantic similarity for "finanspolitisk holdbarhed" with "udgiftspres". A reading of this could be that the GEC can allow themselves a more long-term outlook, and therefore is more concerned about sustainability. On the contrary, DF must be more concerned about rising expenditure. The explanation could be a desire to legitimize budget allocations and/or cuts, while still gaining support for their policy initiatives. The keyword "udgiftspres" is also distant from "finanspolitik" in both graphs. This could emphasize that rising expenditure is an undesirable semantic feature when considering fiscal policy.

Figure 1 - **Finance and Assessment**



The two graphs highlight the semantic properties of our words associated with the themes *finance* and *assessment*. Initially, one can observe interesting features and clusters around the semantics of fiscal policy. For the GEC, assessments such as "ansvar" and "tilbagegang" seems to characterize the semantics around "finanspolitik". This could be explained by the GEC focusing on holding the relevant actors accountable for the consequences of their fiscal policy - hereof "tilbagegang" - if the economy has experienced a recession. Naturally, the GEC also shows "finanspolitisk holdbarhed" as similar to "sunde" to signal the sustainability of a healthy economy or the lack of one. Consequently, the DF has a great distance between "ansvar" and our keywords regarding *finance*. This could potentially symbolize how accountability is not associated with fiscal policy per se. On the contrary, "efterslæb" seems to be similar to "finanspolitik", which could symbolize avoiding accountability by emphasizing a 'passed on' burden from former government coalitions.

Figure 2 - **Finance and Time**

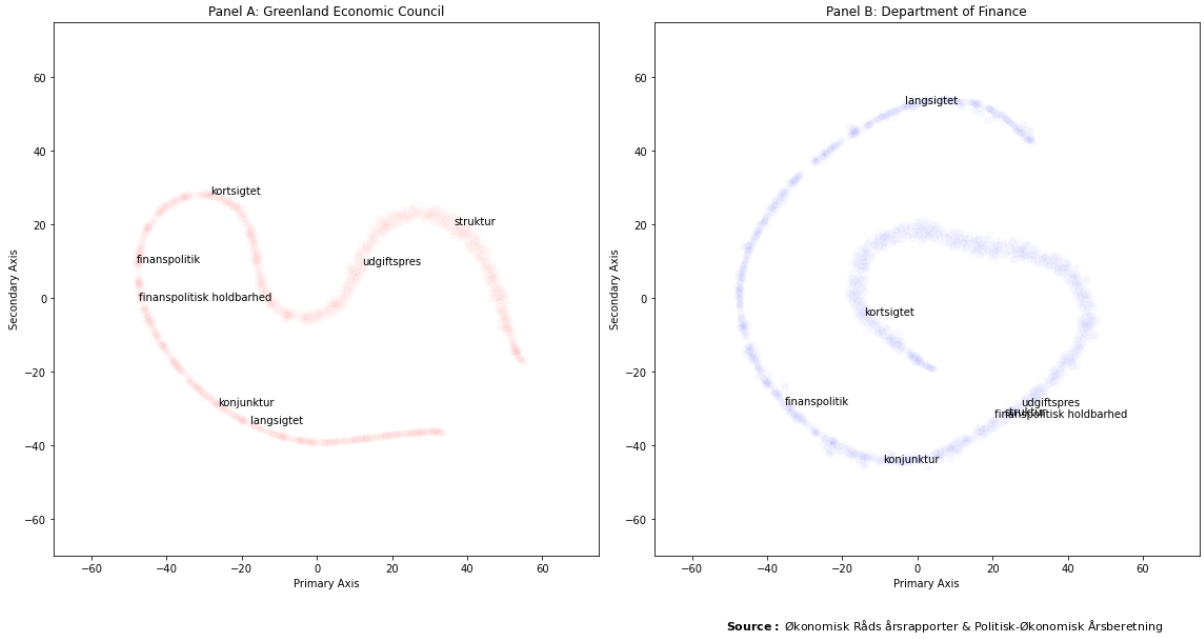


Figure 2 highlights the semantic pattern between the themes *finance* and *time*. It is interesting that "struktur" is closely associated with "finanspolitisk holdbarhed" for the DF (cf. panel B in figure 2). In addition, "konjunktur" is also closer to "finanspolitik" than "finanspolitisk holdbarhed". This may capture a temporal dimension in the way the DF communicates fiscal policy. That is, fiscal sustainability is a structural concept and fiscal policy should address business cycles, rather than structural development. The same pattern is not found for the EC. We interpret this as a break with the classical distinction between business cycles and structural development.

Figure 3 - **Finance and Business**

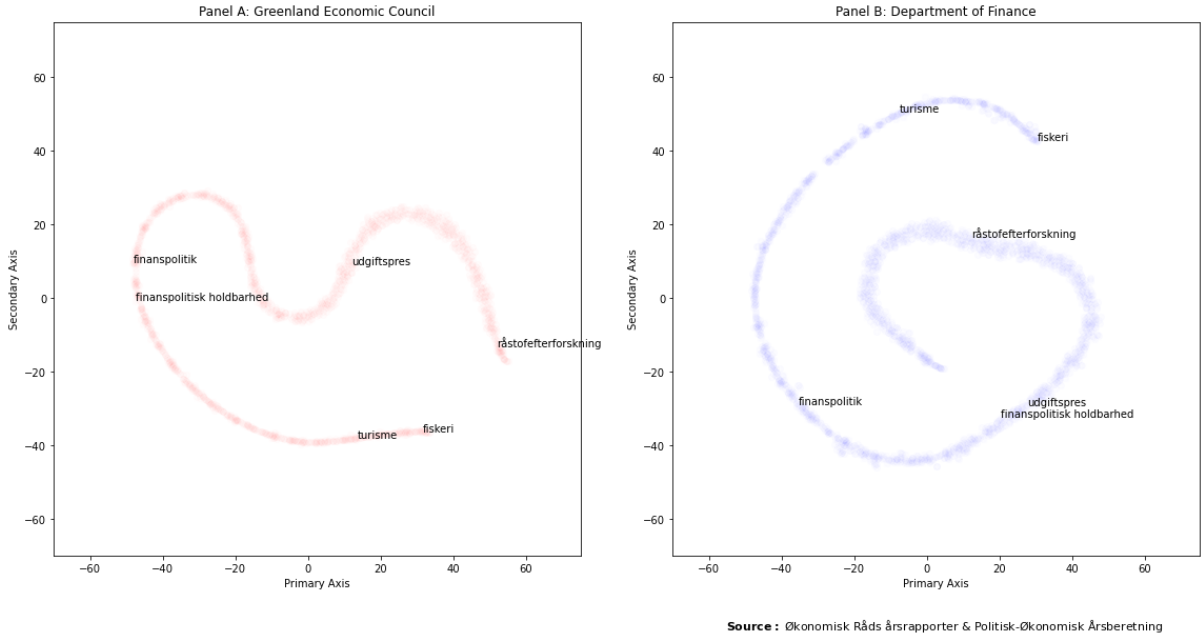


Figure 3 highlights the semantic pattern between the themes *finance* and *business*. It is particularly striking that "råstofefterforskning" is relatively far from "finanspolitisk holdbarhed" for the EC, whereas the opposite is true for the DF. This is probably due to our lack of reports from the EC in the period around 2012-2015. During this time, the mining industry was a hot topic on the political agenda. Besides this lack of data collection, there does not seem to be a close association between the themes as such.

Figure 4 **- Finance and Trends**

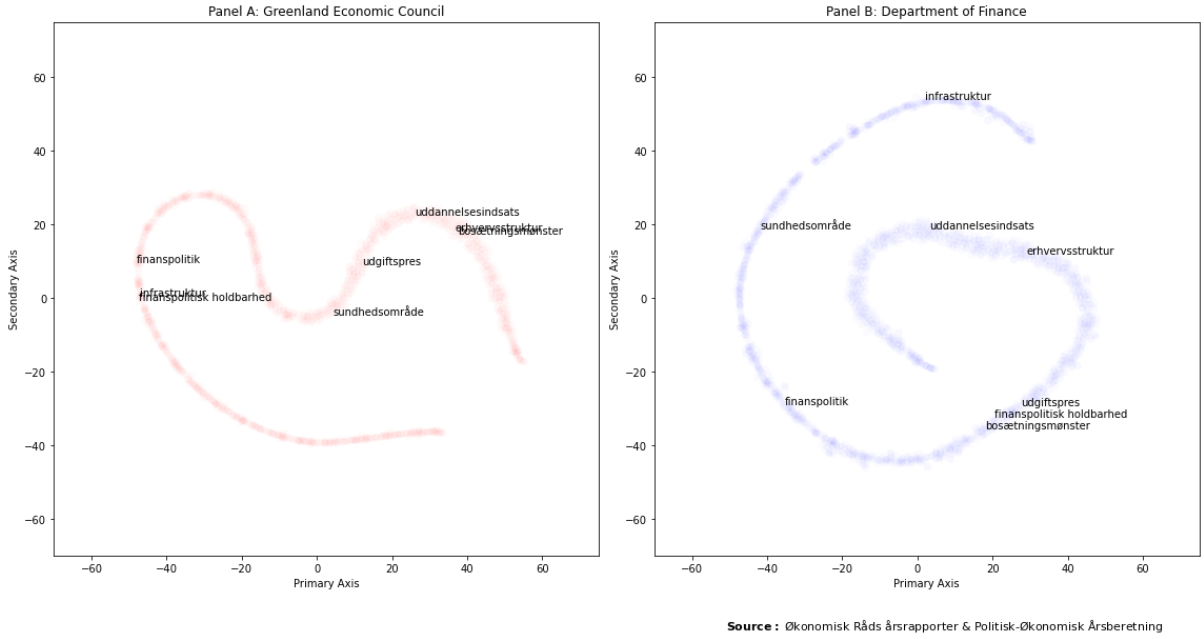


Figure 4 highlights the semantic pattern between the themes *finance* and *trends*. It is interesting that "finanspolitisk holdbarhed" pairs with "infrastruktur" for the EC and "bosætningsmønster" for the DF. This may reflect a certain preference for the assessment of fiscal sustainability. For the EC, "infrastruktur" presumably refers to risk associated with large investments in new national airports (Nationalbanken, 2018). In contrast, "infrastruktur" is almost considered as an outlier for the DF. On the contrary, "bosætningsmønster" may refer to the demographic composition of the population in unviable settlements (Sermitsiaq, 2009).

6. DISCUSSION

This section discusses the limitations of our research. This paper has explored the keyword-level semantic interaction patterns. The approach of the paper is based on the continuous semantic space, which can help grasp the latent semantic relations among the keywords.

In light of the results of this study, several methodological implications can be drawn. In an ideal scenario the paper works with 22 annual reports covering the years 2010-2020. However, due to various complications when reading the data with Py2PDF, only 15 reports were loaded properly. The retrieved reports contain 6 reports from the EC and 9 reports from the DF. In this matter,

the dataset has a larger sample size for the DF. A larger sample size has the obvious advantage of providing more data for this research to work with compared to the small sample size of the EC. Table 1 reports the descriptive statistics for the 15 downloaded reports. It provides separate statistics for the DF and the EC. In terms of numbers of raw words, the DF (143,788) is characterized by a higher number of raw words with respect to the Economic Council (93,286). Numbers of raw words consist of 35.1 pct. more data points for the DF. Table 1 illustrates a comparatively larger number of unique words for the DF (17,179) with respect to the EC (11,487). The DF obtains 33.1 pct. more data points for unique words with respect to the EC. Our findings demonstrate that provision of sample size justifications in the number of raw words, number of words and number of unique words are limited due to working with two different sample sizes. In this case, one can argue our limited sample size has an impact on the empirical findings. Our data is more reliable if we are working with the same sample size of 9 reports for the DF and the EC, but the main reason for not removing 3 reports for the DF from our dataset is the dataset will suffer for working with less data points. In the section 4.4.4 complications of retrieving data, we mentioned that our dataset does not consist of data points from 2012-2015, Greenland has experienced a boom in exploration activities for the mining industry during this period. The economy of Greenland is based on fishing and hunting, but the government had ambitious plans to develop the country's resources in 2010. Greenland's frozen surroundings is one of the world's next great mining frontiers (Mining, 2012). This is another statistical limitation our data suffers from due to the missing data points from 2012 to 2015.

In a previous section on complications of modifying data, we stated that the lemmatizer function has some restrictions. Firstly, our list of selected stop words is incomplete, an alternative was a nltk package's list of Danish stop words but the list of stop words we have selected is larger than the nltk package (262 versus 94 words). The second issue is the lemmatizer applied is a pretrained lemmatizer downloaded from GitHub.

Another drawback of our collection of data is the coronavirus pandemic, which has triggered the deepest economic recession in nearly a century, threatening health, disrupting economic activity, and hurting well-being and jobs (OECD, 2021). This has affected the results of the annual report for 2020. Lastly, our selected themes and the associated keywords have been chosen explicitly by us, and it is not guaranteed that they fully capture our intended purpose. For instance, fiscal

sustainability can be stated and analyzed via many complex paragraphs and combinations of different words. We have a very direct way of measuring the preconcepted semantics, and this may not fully capture the full meaning and context provided in the reports. Our way of deciding upon the words of interest can suffer from reliability issues, it is important to specify that the five themes and the keywords are selected by us, we can consider whether these are the ideal themes for capturing the essentials of semantic syntax.

7. CONCLUSIONS

In this study, we have shown an implementation of Word2Vec for an explorative analysis of 15 public reports of Greenland's economy published by the Department of Finance and the Economic Council of Greenland. Word2vec is applied according to the number of observed data points. The selected pre-trained Word2Vec model is generating fine-tuned word vectors for all words in the corpus. This serves as a novel way of exploring professional reports through visualization of the semantics associated with certain themes and keywords.

Our findings support the idea that the Economic Council has a long run approach to sustainability and the Department of Finance is more concerned about the rising expenditure. The two state actors have a different role in assisting the Ministry of Finance, and this transfers to their semantics. We also find a diverse preference for assessing fiscal sustainability. The Economic Council tends to prioritize trends such as the infrastructure, whereas the Department of Finance is more focused on demographics. It should be noted that our preconception of the subject has a great impact on the analysis. A different preconception would probably have resulted in a different analysis, however, we cannot bypass our preconception in general.

In light of our results, several methodological implications can be drawn. There are 22 annual reports from 2010-2020 due to various complications retrieving the data, we are working with 15 reports. Numbers of raw words consist of 35.1 pct. more data points for the Department of Finance and obtains 33.1 pct. more data points for unique words with respect to the Economic Council. One can argue our limited sample size has an impact on the empirical findings. Finally, it is fundamental to state that the selected themes and the keywords may not assure the essentials of semantic syntax. Overall, the results of this paper appear somewhat successful in visualizing

through Word2Vec. Based on the above mentioned we are not able to draw an unambiguous conclusion on characterizing the semantics associated with economic terminology in economic reports. In theory the economic reports tend to have the identical characteristics, but in our findings, it appears that the Economic Council and the Department of Finance understandably may have different agendas and priorities.

Future studies could potentially examine the semantics of economic concepts over time, rather than the semantics across actors. In the case of Greenland, attitudes towards the economy are influenced by various waves, including the hype around the mining industry, demographic challenges, concerns about healthy ageing and educational disadvantage of the population.

8. REFERENCES

De Økonomiske Råd, 2021. (2012). *Finanspolitisk holdbarhed | De Økonomiske Råd*. 2012. https://dors.dk/offentlige-finanser/finanspolitiske-begreber/finanspolitisk-holdbarhed

Rehurek, Radim (2021). *models.word2vec – Word2vec embeddings — gensim*. 2021. https://radimrehurek.com/gensim/models/word2vec.html

Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., & Sinha, V. (2017). 25 Tweets to Know You: A New Model to Predict Personality with Social Media. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 472–475. https://arxiv.org/abs/1704.05513v1

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 4356–4364. https://arxiv.org/abs/1607.06520v1

Nationalbanken, 2018. *DANMARKS NATIONALBANK ANALYSE 2 8. N O V E M B E R 2 0 1 8-N R. 2 0 GRØNLANDSK ØKONOMI.*

Departementet for Finanser. (2021). *Departementet for Finanser og Indenrigsanliggender - Naalakkersuisut*. https://naalakkersuisut.gl/da/Naalakkersuisut/Departementer/Finans

Finansloven Naalakkersuisut, 2021. (2021). *Finansloven - Naalakkersuisut*. 2021. https://naalakkersuisut.gl/da/Naalakkersuisut/Departementer/Finans/Finanslov

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *3*, 1489–1501. https://arxiv.org/abs/1605.09096v6

Naalakkersuisut Beretning, 2021. (2021). *Politisk Økonomisk Beretning - Naalakkersuisut*. 2021. https://naalakkersuisut.gl/da/Naalakkersuisut/Departementer/Finans/Politisk-Oekonomisk-Beretning

Jens Blom-Hansen, Peter Munk Christensen, Thomas Pallesen og Søren Serritzlew, 2014. *Offentlig forvaltning - et politologisk perspektiv*. Hans Reitzels Forlag, Århus. 1. udgave.

Sermitsiaq. (2009). *D: Luk urentable bygder | Sermitsiaq.AG*. https://sermitsiaq.ag/node/67968

Vágsheyg, K., & Wilhelmsen, Charlotte Sophie, 2016. (2016). *INVESTIGATING GENDER BIAS IN JOB ADVERTISEMENTS WITH WORD EMBEDDINGS MASTER THESIS*.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/JEL.20181020

Le, Mikolov, 2014. (2014). Distributed Representations of Sentences and Documents. *2014*.

Lemmy. (2019). *sorenlind/lemmy: ?Lemmy is a lemmatizer for Danish ?? and Swedish ??* 2019. https://github.com/sorenlind/lemmy

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://arxiv.org/abs/1301.3781v3

Phaseit Inc, & Fenniak, M. (2016). *PyPDF2 Documentation — PyPDF2 1.26.0 documentation*. 2021. https://pythonhosted.org/PyPDF2/

LibHunt, 2021. (2021). *PyPDF2 Alternatives - Python PDF | LibHunt*. 2021. https://python.libhunt.com/pypdf2-alternatives

Råstofeventyret, 2013. (2013). *Råstofeventyret i Grønland: Det handler om statsfinanserne « RÆSON*. 2013. https://www.raeson.dk/2013/rastofeventyret-i-gronland-det-handler-om-statsfinanserne/

Salganik, M. J. (2017) *Bit by Bit: Social Research in the Digital Age*. Open revie. Princeton: Princeton University Press.

Stackoverflow, D. (2021). *Download and save PDF file with Python requests module - Stack Overflow*. 2021. https://stackoverflow.com/questions/34503412/download-and-save-pdf-file-with-python-requests-module

Stopwords, D. (2021). *Dansk stopords liste / Danish stopwords. A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search q*. 2021. https://gist.github.com/berteltorp/0cf8a0c7afea7f25ed754f24cfc2467b

Taddy, 2015. (2015). Document Classification by Inversion of Distributed Language Representations. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, *2*, 45–49. https://arxiv.org/abs/1504.07295v3