Decomposing Text Data: A Case Study of Danish Parliamentary Speeches

Anders S. Olsen[1], Peyman Kor[1]

1 DTU Compute, Technical University of Denmark

Introduction

- **Objective**: Investigate the decomposition method Independent Component Analysis (ICA) on a large corpus of text data.

Decomposition models are used in a variety of data science fields to achieve lower-dimensional internal representations of data sets. In this project, we increase analysis complexity starting from simple term-frequency analyses as the one in Figure 1 to sentiment analysis, and then to the more complex factor model ICA.

- **Data set**: A corpus of parliamentary speeches in the Danish parliament in the years 2009-2016[1].
- **Initial Results**: ICA successfully separates groups of words according to linguistic categories. However, with the setup presented in this project, ICA cannot be used for topic modeling.
- **Conclusion**: ICA can be used to guide linguistic research, however, the sources computed should not be interpreted as unique or absolute.
- **Further Steps**: It could be interesting to use ICA to model linguistic development over the years. Also, interesting linguistic structures may be revealed when increasing the number of independent components. For topic modeling, other methods such as Latent Dirichlet Allocation (LDA) could be investigated.

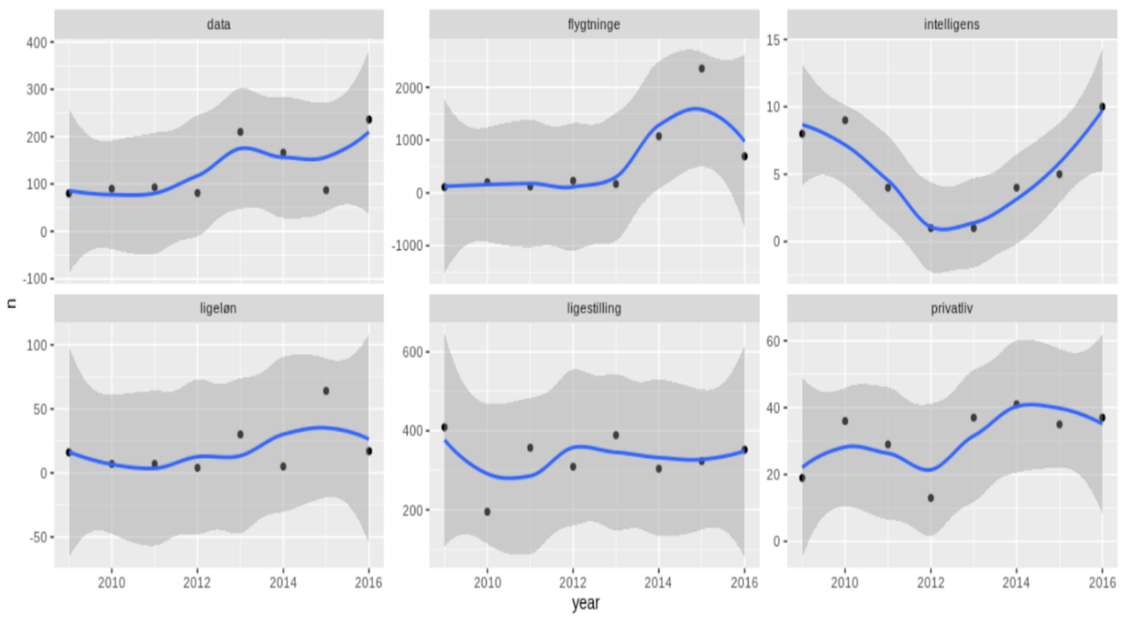

Figure 1: Evolution of key-words spoken in the Danish parliament.

Data preprocessing and cleaning

As data were provided in XML-files, data extraction is the natural first step. In R, the subbranches "Navn", "Rolle", "Starttid", "sluttid", "tekst" of the speeches were extracted. The final data was aggregated in the table with around 400,000 speeches in the Danish Parliament in the years 2009-2017. Thereafter, to do text analysis, the punctuation marks and numbers were removed as well as other symbols. This resulted in a table of around 40 million words in total. However, a list of common and non-significant words were removed from the data to facilitate more insightful analysis of the words. A list of stop-words[2] was used to filter out words with no sentimental meaning, reducing the amount of analyzed words to 16 million. We added common danish abbreviations as well as party name abbreviations to this list.

Descriptive Analysis

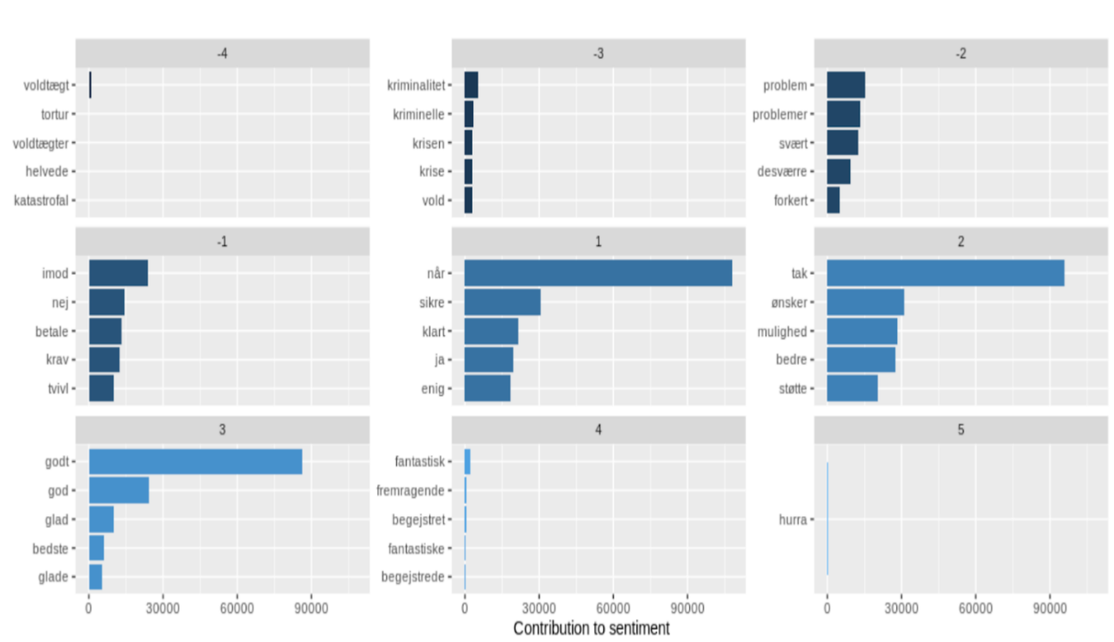

Figure 2: Sentiment analysis of words spoken in the parliament

Figure 2 shows the most frequently used words according to their sentiment, as provided in [3]. This shows that generally more positive words were used than negative, which may be reflected in the very formal tone used in the parliament.

Independent Component Analysis

As a member of the family of factor models, ICA decomposes a matrix $\mathbf{X}$ into a mixing matrix $\mathbf{B}$ and a source matrix $\mathbf{S}$ according to

$$\mathbf{X} \approx \mathbf{BS} \tag{1}$$

Whereas PCA computes uncorrelated sources, ICA aims at producing sources that are statistically independent. For Gaussian data, uncorrelatedness implies independence. Thus, ICA computes sources that are non-Gaussian. In this project, this is achieved by maximizing Kurtosis between sources.
Inspired by [4], the data matrix $\mathbf{X}$ is created by extracting the $N = M = 1000$ most frequent words from the pool of text. For each word w_i, $i = 1..N$ the number of occurrences of word w_j, $j = 1..M$ either just before or just after word w_i is noted at position X_{ji} (Figure 3). To dampen the differences between word-context frequencies, the log of the matrix plus one was taken: $\log(\mathbf{X} + 1)$

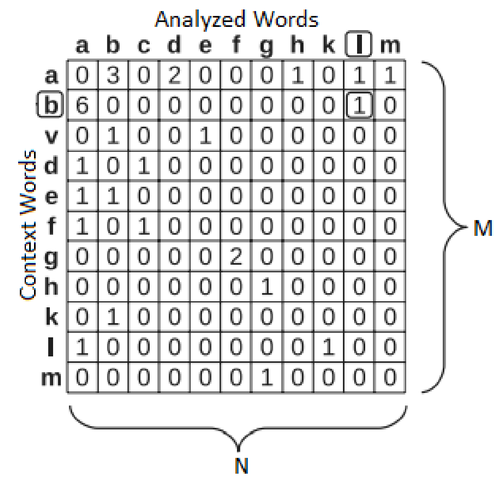

Figure 3: The word-context matrix $\mathbf{X}$ used for ICA[1]. In the word pool, *b* appears immediately before or after *l.* which is noted in the matrix.

To use ICA, the data need to be pre-whitened. This was achieved by running applying an SVD-decomposition: $\mathbf{X} = \mathbf{UDV}^T$. Then, $\mathbf{V}^T$ was used for further analysis, as depicted in Figure 4.

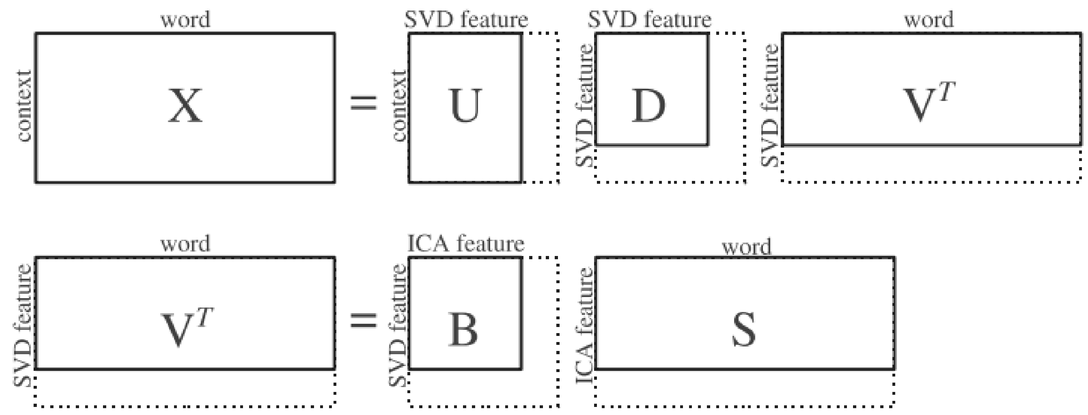

Figure 4: The pre-whitening procedure prior to ICA.

Results

Source 1	Source 2	Source 3	Source 4	Source 5
lang	ganske	ministerens	dansk	sidste
lange	utrolig	ordførerens	socialistisk	første
års	relativt	korte	konservative	enkelt
stykke	rimelig	regeringens	danske	tredje
taget	ret	venstres	socialdemokratiske	kommende

Table 1: The 5 words with the largest representation in each of 5 components.

Table 1 shows the first 5 out of $c = 10$ computed independent components. Words in each source are sorted by their absolute contribution.

Discussion

- Results from ICA should not be interpreted as absolute. To develop linguistic categories on the danish language, text contributions from a wider mixture of people are needed. Even then, results should only be interpreted as guidelines to nudge research in specific directions.
- In this project, we did not investigate the number of independent components c needed to represent the data set. Utilizing the fact that components split when c increases could reveal information about the link between categories.

References

[1] Dorte Haltrup Hansen. The danish parliament corpus 2009 - 2017, v1, 2018. URL http://hdl.handle.net/20.500.12115/8. CLARIN-DK-UCPH Centre Repository.

[2] Bertel Torp. Danish stop-words, 2020. URL https://gist.github.com/berteltorp/0cf8a0c7afea7f25ed754f24cfc2467b.

[3] Finn Aarup Nielsen. Danish word sentiments, 2020. URL https://raw.githubusercontent.com/fnielsen/afinn/master/afinn/data/AFINN-da-32.txt.

[4] Timo Honkela, Aapo HyvÄrinen, and Jaakko j. VÄyrynen. Wordica: Emergence of linguistic representations for words by independent component analysis. *Nat. Lang. Eng.*, 16(3):277–308, July 2010. ISSN 1351-3249. doi: 10.1017/S1351324910000057. URL https://doi.org/10.1017/S1351324910000057.

s154043@student.dtu.dk s191828@student.dtu.dk