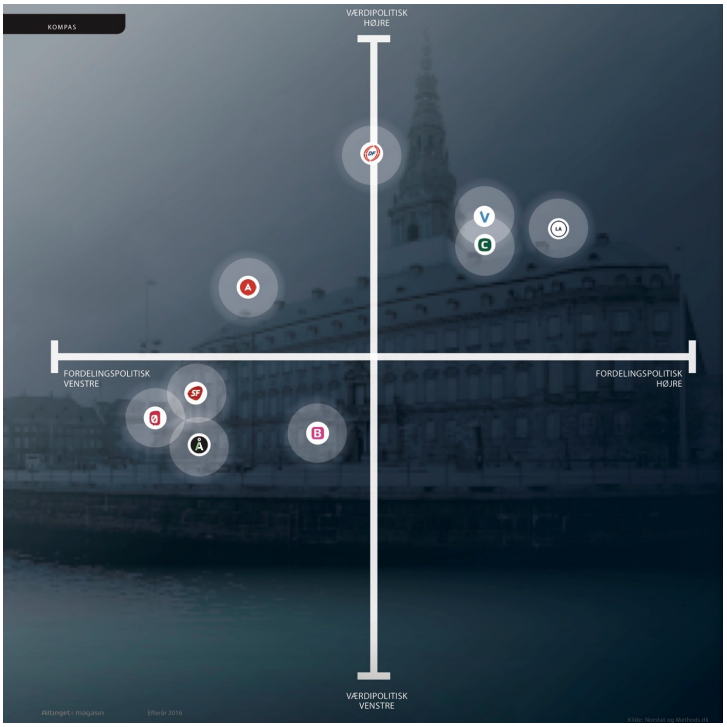
Danmarks Tekniske Universitet

Andreas Nilausen, s183915
August Semrau Andersen, s183918
William Marstrand, s183921

# Signals and Data
# 02462

## POLITICAL PARTY CLUSTERING




May 10, 2019, Danmarks Tekniske Universitet.

## Abstract

Natural Language Processing models are getting increasingly complicated with companies like Google using state-of-the-art NLP technology. Still, English is the dominant language and languages such as Danish are not getting a lot of attention.
This paper studies text data from political statements about key issues found on the websites of Danish parliamentary parties to try and group the different parties in accordance to the political landscape. The study performs latent sentiment analysis using a simple unigram Bag-of-Words model and Principle Component Analysis for dimensionality reduction. The Euclidian distances between the center of masses for the individual parties are determined and used to decide how close they are to each other. Statements about the European Union, immigration politics, and other popular issues are investigated and the effect of document splitting is examined. It is found that splitting the documents into smaller corpus sizes is optimal and patterns of clustering are found to mostly agree with the model of Altinget[1].
Further research on this topic is needed and could include stronger statistical analysis and an optimzed data preprocessing.

# Contents

# 1 Introduction

Natural Language Processing programs have emerged in different languages, though English having massive data resources available is still dominating the field. But, other languages spoken by fewer people, are becoming relevant for companies such as Apple and Google with services like Siri and Google Assistant and for governments trying to keep up with digitization. An example of this is the Danish language which proposes an interesting case to study.

## Research Question

With this paper we want to investigate the possibility of mapping the political landscape of Danish political parties through *latent semantic analysis* (LSA). We perform LSA with a unigram Bag-of-Words (BoW) on textual descriptions of political key issues for the parties using *Principal Component Analysis* (PCA) to reduce the vector space.
We expect parties that share similar political beliefs to use similar rhetorical strategies, thereby being grouped together in the LSA.

# 2 Method

The experiment consists of LSA performed on text data from different parties using a unigram BoW with PCA in $k$ dimensions, adjusting $k$ to get the most promising results.

## 2.1 Data Collection and preprocessing

The text data is mined from the parties' websites. Opinions on the following political issues has been scraped: *climate*, *immigration*, *law*, *education*, *healthcare*, *European Union*, and *taxes*.
The preprocessing of the texts includes stemming using *SnowballStemmer* from the library *nltk.stem.snowball* with the Danish setting, and filtering on stemmed stopwords[1].

Four different corpus settings were used:

1) A single document for each party containing their EU politics
2) A single document for each party containing their immigration politics
3) Each political key issue as a single document creating 7 document vectors per party
4) Political key issues split up into 85 lines, splitting on new lines, giving 85 document vectors per party.

## 2.2 Experiment Implementation

The experiment is implemented using Python 3.6.7. The source code can be found on Gitlab.[3]

### 2.2.1 Bag of Words

The unigram BoW is implemented as a numerical matrix representation of the word count in each party document. A row representing a word and the columns representing the different party documents. The vocabulary is the set of words from all documents after filtering out stop words. The word count is done by checking how many times every word in the vocabulary appears in each party text. This method of word classification is simple as it only takes into account how many times an individual word is present (*unigram words*), and not the order with which the words appear.

### 2.2.2 Principle Component Analysis

PCA is used for feature extraction. It compresses the BoW by reducing the dimensions in the vector space to the axis with highest variance. We subtract the mean of each variable and thereby center the data around the origin. Then the covariance matrix and it's eigenvectors and eigenvalues are calculated. The normalized orthogonal eigenvectors are then used to form a new basis fitted to the BoW data.

The extracted features of each document (each party's political key issues) are visually represented as points in a $n$-dimensional vector space, wherein differences between the documents can be understood as the distance between vectors and therefore the spatial relations between them.
This creates vectors for each party, of which the center of mass is computed resulting in a single

[1] See Appendix 5.2 for list of stemmed stop words

mean vector for each party, representing their overall use of words as a projected point in the PCA vector space. The euclidean distances (equation 1) between these centers of masses are then calculated and used for evaluating how far apart the different parties are from each other.

$$Dist(\boldsymbol{p}, \boldsymbol{q}) = \text{Dist}(\boldsymbol{q}, \boldsymbol{p}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{1}$$

# 3 Results

Below is a table mapping the political parties and their party letter:

| **Party Name** | **Party Letter** |
|---|---|
| Enhedslisten | Ø |
| Socialistisk Folkteparti | SF |
| Socialdemokratiet | A |
| Radikale Venstre | B |
| Alternativet | Å |
| Venstre | V |
| Liberal Alliance | LA |
| Det Konservative Folkeparti | C |
| Dansk Folkeparti | DF |

In the first two tests of the experiment, we ran LSA on one document per party, the first test being their political standpoints in the EU and the second being their standpoints of immigration politics. Looking at the plots and reviewing the distances between parties the clustering of parties did not match the expected grouping from the Danish political compas. For tables and 2D plots showing these, see page appendix 5.1.

In the third test, each party had 7 documents consisting of the same political issues. Here the results did not match the expected grouping either. For tables and 2D plots see appendix 5.1.3.

In the fourth and final test, we split the 7 documents for each party up into 85 lines, meaning each of the 9 parties had 85 documents. The distances between the parties were calculated in $k = 3$ dimensional space which was found to be the optimal number of dimensions for grouping of the parties. The test found a grouping much in line with the general perception of the parties political relationships from the Danish political compas.
The parties V, LA and C (right-wing parties) clustered together, parties A, Ø and SF (left-wing parties) clustered together, while DF is placed in between these two groupings. A table showing the nearest and most distant party for each party can be seen below, together with a 2D PCA visualization.

**Popular Topics (85 lines)**

| | Ø | SF | A | B | Å | V | LA | C | DF |
|---|---|---|---|---|---|---|---|---|---|
| **Nearest Party** | SF | Ø | DF | Ø | B | LA | C | LA | C |
| **Most Distant Party** | Å | Å | Å | A | A | Å | Å | Å | Å |

Table 1: Popular Topics Politics (docN = 85)

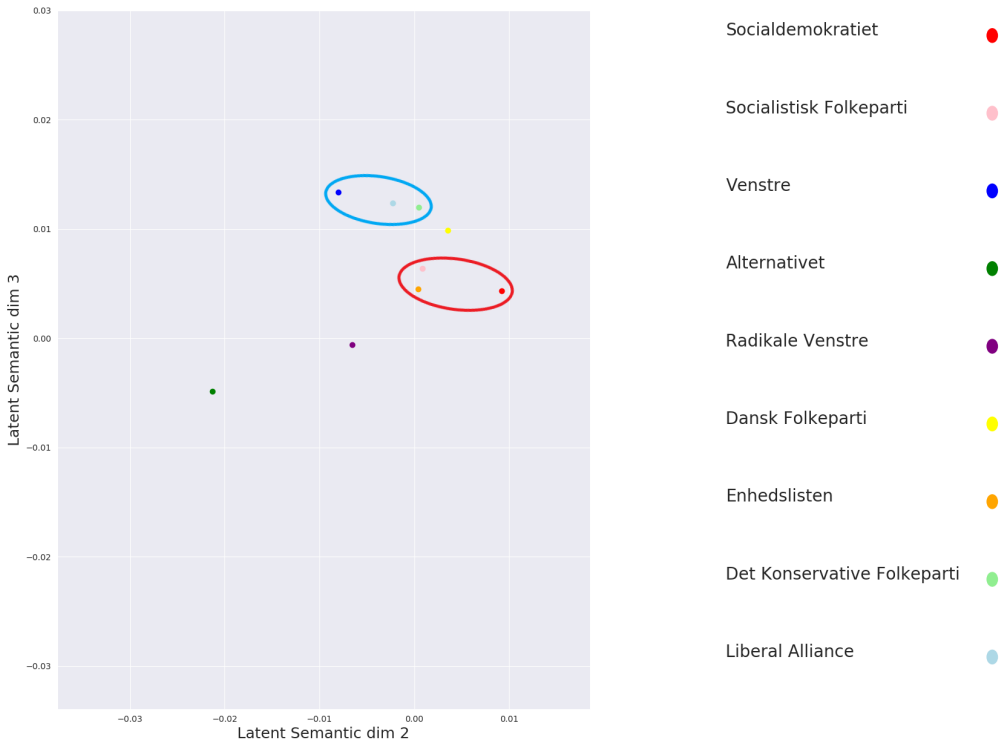



Figure 1: 85-doc per party, 2 Dimensional PCA Visualisation

### Heat Table of Popular Issues Distances (85 lines)

The heat map in table 2 is made from experiment setup 4 with 85 lines as documents per party. It shows each party and their distance in relation to the other parties
The heatmap is shown with red colors representing relatively large distances from the given party and green colors representing smaller distances to the party.

| | Ø | SF | A | B | Å | V | LA | C | DF |
|---|---|---|---|---|---|---|---|---|---|
| Ø | | 0,0019 | 0,0088 | 0,0086 | 0,0236 | 0,0122 | 0,0083 | 0,0075 | 0,0062 |
| SF | 0,0019 | | 0,0086 | 0,0102 | 0,0248 | 0,0113 | 0,0068 | 0,0056 | 0,0043 |
| A | 0,0088 | 0,0086 | | 0,0165 | 0,0318 | 0,0194 | 0,0140 | 0,0116 | 0,0079 |
| B | 0,0086 | 0,0102 | 0,0165 | | 0,0153 | 0,0140 | 0,0137 | 0,0144 | 0,0145 |
| Å | 0,0236 | 0,0248 | 0,0318 | 0,0153 | | 0,0225 | 0,0257 | 0,0275 | 0,0288 |
| V | 0,0122 | 0,0113 | 0,0194 | 0,0140 | 0,0225 | | 0,0058 | 0,0086 | 0,0121 |
| LA | 0,0083 | 0,0068 | 0,0140 | 0,0137 | 0,0257 | 0,0058 | | 0,0028 | 0,0063 |
| C | 0,0075 | 0,0056 | 0,0116 | 0,0144 | 0,0275 | 0,0086 | 0,0028 | | 0,0037 |
| DF | 0,0062 | 0,0043 | 0,0079 | 0,0145 | 0,0288 | 0,0121 | 0,0063 | 0,0037 | |

Table 2: Heat Map in continuation of table of 2 Dimensional PCA

# 4 Discussion

Based on the results in the fourth setup of the experiment, it seems possible that a grouping of the Danish political parties according to their political beliefs can be done based on their usage of words. Still, further studies are needed.

## 4.1 Considerations

During the experiment we found that the choice of data and preprocessing had a big impact on the outcomes. Firstly, the same texts on political key issues were chosen for all parties. This was done to avoid bias leading to parties being evaluated as closer or further from each other simply because they were being evaluated on different political issues. An example of this was seen in the pre-experimental tests where only two of the nine parties had documents on politics about children, and therefore had much shorter Euclidian distance between them than expected.

Secondly, only using documents on issues from the parties' websites limited the size of the dataset. It would be interesting to use all transcripts from all discussions in the parliamentary hall *Folketinget*, published by DR [2]. On the other hand, using data on political key issues specific to certain parties may result in greater variance in the dataset as parties often differ in opinions on popular issues, especially at the time of an upcoming election. Furthermore, due to time restrictions the statistical work on the empirical results was very limited. Making further experiments a necessity to reject or accept the hypothesis.

Thirdly, filtering of the datasets, especially the removal of stop words, could be optimized to reduce noise in the data and potentially result in a clearer LSA-analysis.
Our analysis of party distances is based on the Euclidian distance formula as the length of the straight line between two points in $n$ dimensions. Another method of evaluating patterns in the data could be using the cosine similarity between the angles of each of the parties' eigenvectors in the PCA. This method of differentiation takes both the lenght and the angle between the eigenvectors into account.

With these considerations in mind future research on this topic would be very interesting. Moreover using $n$-grams could be beneficial and help to find contextual relations between multiple words, possibly finding even greater differences between the parties.

As we are writing this paper, the Danish parliamentary election has begun and many Danish citizens will over the next months have to decide on which party's political agenda they agree with the most. In this process, voters will be prone to change their opinions based strongly on what they read in the news and especially on social media. Having a tool based on this and further experiments might help categorising the different parties and evaluating what topics are closest to each party which could be a great help for voters.

# References

[1] Lasse lange
Her er det nye politiske kompas
https://www.altinget.dk/artikel/det-nye-politiske-kompas

[2] Kåre Kildall Rysgaard
Klima, sex og cigaretter: Tjek, hvilke politikere der taler om det, der er vigtig for dig
https://www.dr.dk/nyheder/politik/folketingsvalg/klima-sex-og-cigaretter-tjek-hvilke-politikere-der-taler-om-det-der

[3] Gitlab - Source Code
https://gitlab.gbar.dtu.dk/s183921/pca-project

# 5 Appendix

## 5.1 Nearest Parties

Below are results from the PCA distance evaluations described in tables. Each table is labeled with the investigated policy and names of the parties are listed in the table.

### 5.1.1 EU Politics

| | Ø | SF | A | B | Å | V | LA | C | DF |
|---|---|---|---|---|---|---|---|---|---|
| **Nearest Party** | DF | B | DF | V | DF | B | DF | DF | B |
| **Most Distant Party** | C | Ø | Ø | Ø | Ø | Ø | SF | Ø | Ø |

Table 3: European Union Politics

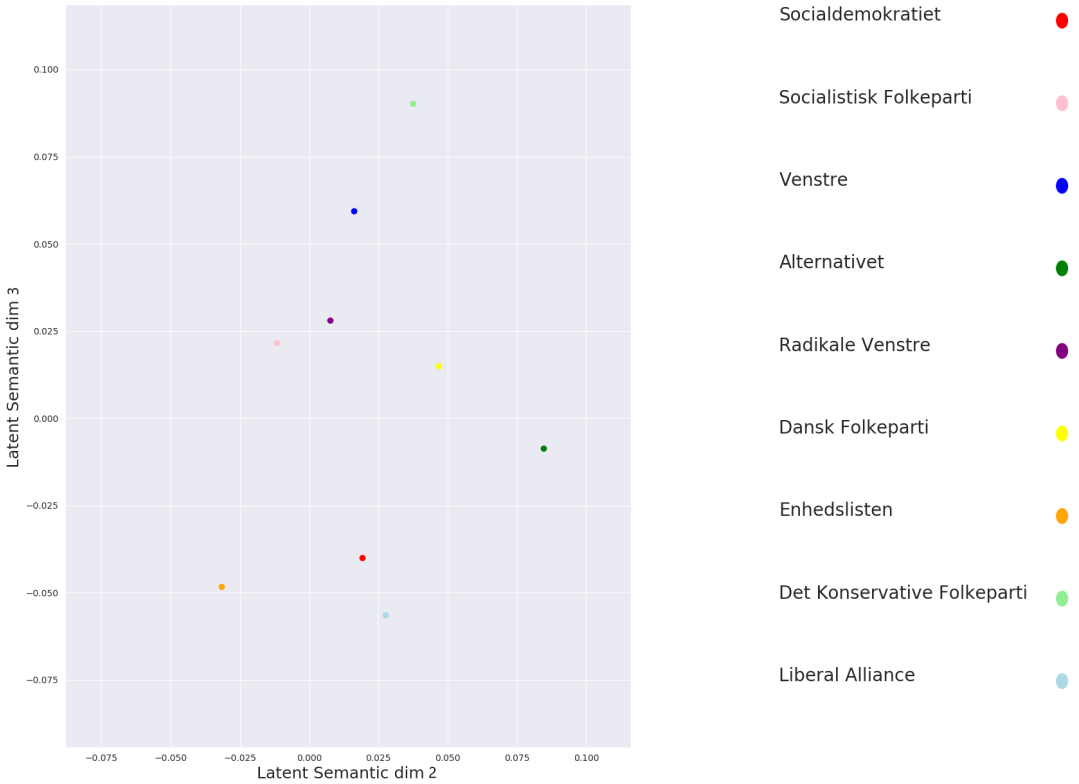



Figure 2: 2D PCA Visualisation of EU-Politics

### 5.1.2 Immigration Politics

| | Ø | SF | A | B | Å | V | LA | C | DF |
|---|---|---|---|---|---|---|---|---|---|
| **Nearest Party** | LA | LA | DF | LA | B | LA | DF | LA | LA |
| **Most Distant Party** | V | Å | Å | V | V | Å | Å | V | Å |

Table 4: Immigration Politics

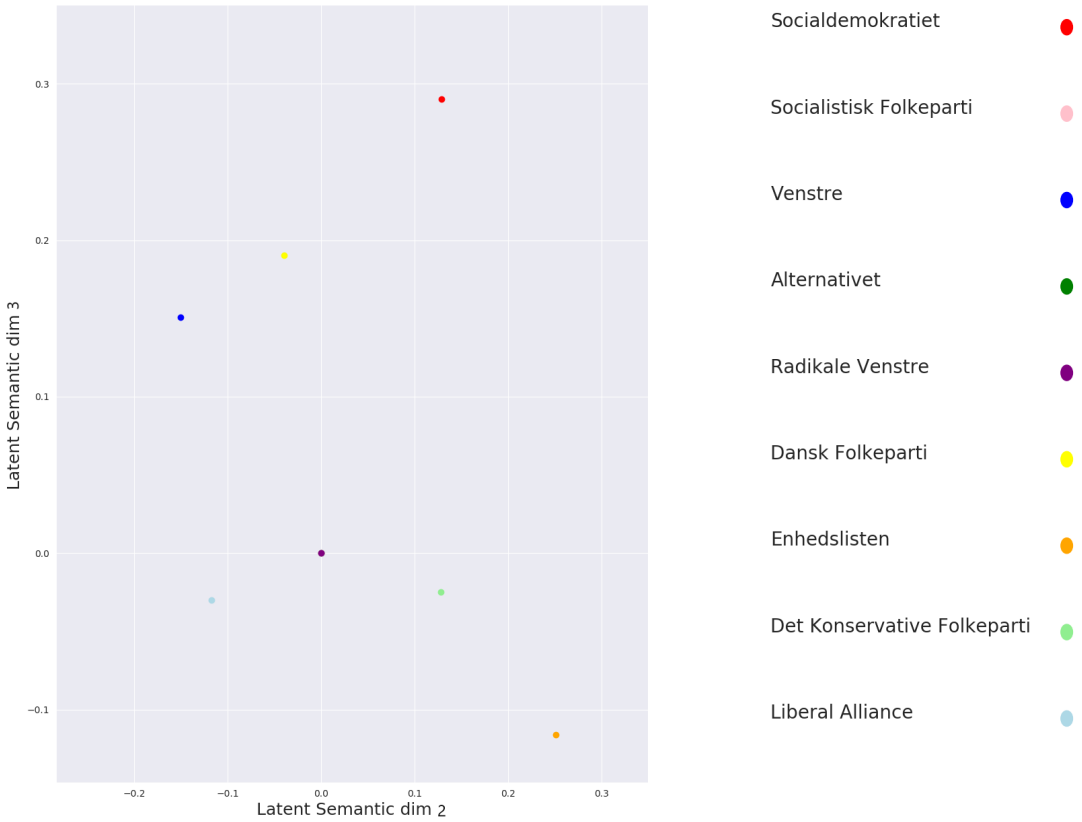



Figure 3: 2D PCA Visualisation of Immigration-Politics

### 5.1.3 Popular Topics (full documents)

| | Ø | SF | A | B | Å | V | LA | C | DF |
|---|---|---|---|---|---|---|---|---|---|
| **Nearest Party** | SF | Ø | SF | C | Ø | SF | DF | B | LA |
| **Most Distant Party** | A | A | Å | Å | A | A | A | A | A |

Table 5: Popular Topic Politics for full documents

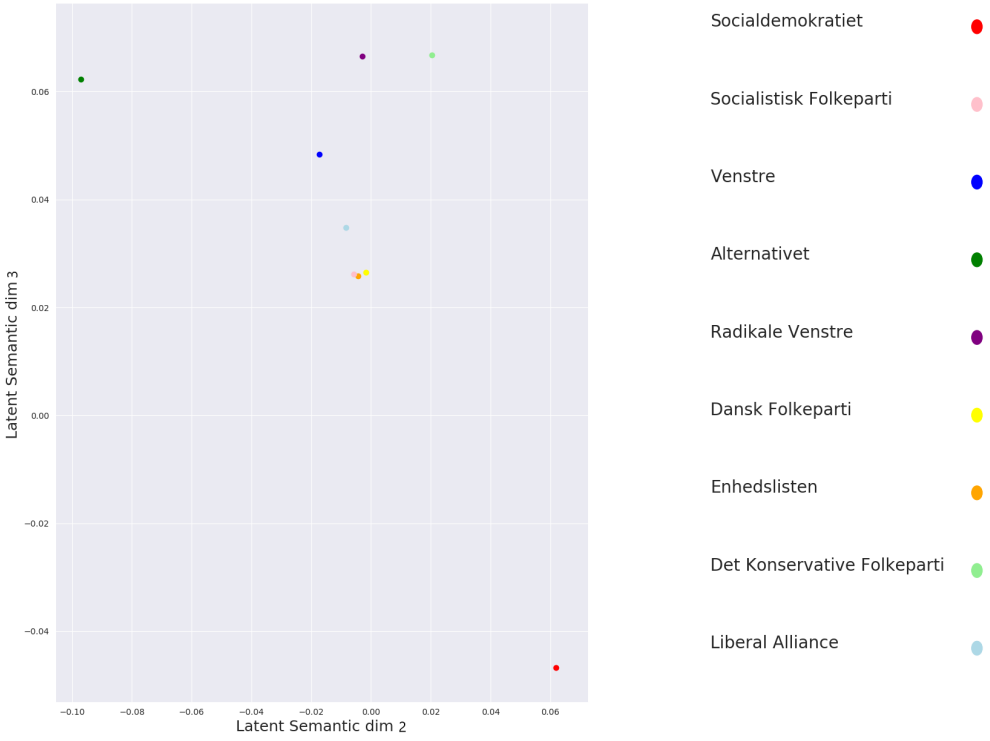



Figure 4: 2D PCA Visualisation of Popular Topics as entire corpus

## 5.2 List of Stop Words - Stemmed using SnowballStemmer

ad af aldr all alt and and andr at bar beg blev bliv bliv da de dem den den der der det det dig din din dis dit dog du eft ej ell en end ene enest enhv er et far fem fik fir fler flest for fordi for fra få får før ham han han har havd hav hej helt hend hend her hos hun hvad hvem hver hvilk hvis hvor hvordan hvorfor hvornår i ikk ind ing int ja jeg jer jer jo kan kom kom kom kun kun lad lav lidt lig lil man mand mang med meg men men mer mig min min mit mod må ned nej ni nog nog nogl nu ny nyt når nær næst næst og også om op os ott over på se sek selv ser ses sig sig sin sin sit skal skul som stor stor syv så sådan tag tag thi ti til to tre ud und var ved vi vil vil vor vor vær vær alen all alligevel altid bag bland burd bør den dereft derfor derfra deri dermed derpå derved egen ell endnu ens ent flest foran først gennem gjord gjort gør gør gør hel hel hen henov hereft heri hermed herpå hvilk hvilk hvoreft hvorfra hvorh hvori hvorimod hvorved igen igennem imellem imen imod indtil lang lav lav ligesom læng mellem mest mindr mindst måsk nem nogensind nok omkring overalt sam sam selvom sen sid stad syn synt såled tem tid tilbag tit uden udov undtag via vor vær øvr