Master of Science Thesis

Econometrics & Management Science

Using correspondence analysis to map the portrayal of Dutch political parties in newspapers

March 20, 2014

Author:
Theun van Vliet
Student number:
303407

Supervisor:
Dr. Michel van de Velden
Co-reader:
Prof. dr. Philip Hans Franses

Abstract

In this thesis, the portrayal of Dutch political parties by Dutch national newspapers is analyzed. Relevant terms related to politics are found and associated with newspapers, political parties and over time. This is done by finding the frequencies of occurrences in each newspaper, co-occurrences with political parties and occurrences within each time period. Using correspondence analysis (CA) these frequencies are visualized. In addition, using Procrustes rotation the CA solutions are aligned to each other to examine the movements over time. With bootstrapping techniques, stability measures are calculated and confidence ellipses found to be included in the visualization. The extracted relevant terms are found to be associated differently with each political party. The portrayal shows that a division exists between coalition and opposition parties, both for co-occurrences of terms and within newspapers. Moreover, a division is found between financial and green, humanistic themes among which movements of parties are shown.

Keywords: Correspondence Analysis; Text analysis; Procrustes rotation; Bootstrapping; Political science

I would like to thank my parents for their continuing support. Philip Hans Franses for introducing me to Econometrics. And last but not least my thesis supervisor Michel van de Velden for his insights and guidance.

Contents

1 Introduction **2**

2 Literature review **4**

3 Methodology **7**

3.1 Preparing the data . . . 7

3.2 Extracting relevant terms . . . 8

3.3 Correspondence Analysis . . . 9

3.3.1 Correspondence Matrix . . . 9

3.3.2 Singular Value Decomposition . . . 10

3.3.3 Calculation of Coordinates . . . 10

3.3.4 Interpretations . . . 12

3.3.5 Contributions . . . 12

3.3.6 Supplementary points . . . 13

3.3.7 Bootstrapping . . . 13

3.3.8 Frequency Matrices . . . 15

3.4 Procrustes Rotation . . . 16

3.4.1 Procedure . . . 16

3.4.2 CA Contributions After Rotation . . . 17

3.4.3 Congruence Coefficient . . . 18

4 Results **19**

4.1 Data . . . 19

4.2 Extracting relevant terms . . . 20

4.3 Correspondence Analysis . . . 21

4.3.1 Parties by newspapers . . . 21

4.3.2 Terms by time period . . . 23

4.3.3 Terms by parties over the whole period . . . 25

4.3.4 Terms by parties per quarter . . . 26

5 Conclusion **30**

A Dutch stopwords **34**

B Extracted terms for each period **35**

C Python script for extracting articles from LexisNexis article list **36**

D Co-occurences per quarter **40**

1 Introduction

A democracy requires that people are well-informed about the political system. Therefore, news media have the important job to give a clear overview of the political parties. However, the dynamics of politics can still be difficult to comprehend. In the last decade, the electorate's voting behavior in the Netherlands has been swinging a lot. There have been five elections for the House of Representatives in 11 years, where both right- and left-wing parties have won and lost multiple times. In 2010, the PVV (Party for Freedom) obtained a large part of the vote, while it only existed 5 years (NOS, 2010). Similarly, the SP (Socialist Party) has shown an unexpected popularity amongst Dutch voters in 2012 (Volkskrant, 2012).

The media, particularly television stations and newspapers, are an important news source for many people. Therefore they play a large role in the formation of public opinion, whether it is by supplying information or through examining voting behaviour. So how do these media portray the political parties and its relationships to each other? Moreover, how do these relationships evolve over time and what topics are important? And what is the influence of the media on these relationships?

In my thesis I address these questions focusing specifically on the terms used in national newspapers. I give insights into how newspapers portray the Dutch political parties and answer questions on the course of events that happened over the past year. These insights are relevant for political parties and newspapers alike. For parties, when they understand the relationships and the way they are portrayed, they could finetune their media strategy. They could uncover the news items and topics that are important to their exposure and the ones that are not. As for newspapers, their objectivity is an important factor for their credibility. This can be researched by uncovering any associations newspapers contribute with certain political parties. Moreover, this can be interesting for all Dutch voters. The method provides an overview of the important news items and how parties relate to these items. This allows the information to be much more easily digested and used to decide which party to vote on.

Before doing the analyses I have phrased the following research question:

How do Dutch national newspapers portray the different Dutch political parties and their relationships?

Within this research question, I have stated the following sub-questions to further specify my research and answer the broader main question:

- Can we identify (sets of) terms that are particularly relevant for different political parties?
- Can we describe relationships between political parties and the relevant words mentioned in newspapers? And if so, do these relationships change over time and between periods?
- Do relationships and/or portrayals consistently differ for different newspapers?

These sub-questions provide an in-depth analysis into the relationships of terms, parties, time periods and newspapers. In my research, I examine the relevant politically related terms in newspapers over a time period. I map the relationships between these terms, incorporate political parties, and examine the changes of these relationships over time and between newspapers.

I select the relevant terms using the reference text technique by Zuell (2008). It discovers terms related to a notable event within a certain period. Counting the number of occurrences of these terms in each article, I then summarize them into a frequency table. This provides the number of mentions of terms in certain contexts, namely newspapers and time periods. To visualize this, I use correspondence analysis (CA) (Greenacre, 1993). CA finds associations between the rows and columns of the frequency table. This is done by reducing the dimensionality of the table - usually to two dimensions - to summarize the main associations. It visualizes both the columns and rows as points within these dimensions. The relative positions between these points show the associations between the rows and columns of the frequency table.

I apply correspondence analysis in three ways:

- Using term and party frequencies in newspapers
- Using term and party frequencies over certain time periods
- Using parties and terms co-occurrences within articles

Using these three perspectives, I answer the stated research question from different angles. First, which newspapers portray which political parties. Second, what terms are used in which time period and what is the relationship with each political party over time. Lastly, which terms are associated with which political parties.

By visualizing the relationships between terms and political parties in newspapers, this constitutes a way to describe the Dutch political landscape. This portrayal is dynamic over time. As different events happen or the focus on certain topics changes, the different stances of parties on these different matters also change. By doing CA for each time period, I incorporate these dynamics to explain the movements over time.

The thesis is structured in the following way: In chapter 2, I provide a literature review on methods to analyze texts using quantitative methods, including correspondence analysis. In chapter 3 the methodology is explained. Here I go into detail describing the extraction of relevant terms, how these are visualized using CA and rotated over time to improve interpretability. The results of these methods are described in chapter 4, which I will discuss in the conclusion in chapter 5.

2 Literature review

In the past, Dutch political news has been analyzed in several occasions. For example, Kleinnijenhuis et al. (2003) have done research on how political parties are portrayed in the Dutch news and how this influenced the election results in 2002. They used their own so-called NET method which is a network text analysis tool, where each article's headlines or text had to be coded with actors, issues and standpoints. This gave them an overview of the parties' stance towards each other and towards political topics. It provided a better interpretation and explanation of the events that happened during the 2002 election. Just like this research, this thesis intends to explain the interaction of political parties and their stance towards topics using the reports of newspapers.

Grijzenhout et al. (2010) applied opinion mining techniques to extract the paragraphs in Dutch Parliament speeches that contain an opinion. This was done by matching the terms in the speeches with a list of terms used to express an opinion. They compared two models of finding an opinion. One model counted the number of subjective words in a paragraph. The other model only checked whether a subjective word was present in a paragraph or not. Three machine learning methods were used to predict these values and their accuracies were compared. In addition, by matching the terms with a list of terms tagged as having a positive or negative association, they made a score for the position of the paragraphs with regard to the corresponding topic. With the three machine learning methods they predicted whether the speaker of the speech was for or against the opinion expressed in that paragraph. In both of the cases they were successful, using Naive Bayes classifiers and Support Vector Machines.

Social media has also been used to analyze political sentiment and moreover predict election results, for example from the online social media platform Twitter. Bermingham and Smeaton (2011) examined Irish election results and found messages posted on Twitter that were related to the Irish political parties. They considered the number of occurrences of a party and the sentiment towards a party separately. Then they used linear regression to build a predictive model of the election results. Their findings concluded that the number of occurrences is a better predictor than the sentiment of a message on Twitter.

Conover and Gonçalves (2011) also used Twitter for research on political sentiment. They predicted political alignment of users on the service. First, they did an analysis on the content of the messages, using the mentioned hashtags (user-defined labels in a message). In addition, they used the network of the users as predictors. In both cases, they applied Linear Support Vector Machines to predict whether a user was right or left-wing in the political spectrum. Both the content and network analysis were very effective in prediction, with accuracies above 90%, the network analysis being marginally more accurate.

Similarly using other web documents, like weblogs, Efron (2005) was able to predict political alignment of these documents. This was done by considering the cocitation of the documents. Efron looked at the links between web documents by examining where the hyperlinks in a document lead to. The idea is that documents with the same left- or right-wing orientation have

similar hyperlinks. He used a simple probabilistic model to measure the association between documents. Using this association he could predict the political orientation with an accuracy higher than 90%.

The above-mentioned articles focus on the prediction of political alignment and sentiment using terms and relations. In my thesis I visualize the relationships and associations of terms and political parties occurring in newspapers. The reason for choosing newspaper articles is that they are very well suited for text analysis, since in general they are structured and annotated. This in contrast to e.g. spoken text or text from social media. Therefore, a lot of text analysis has been applied to newspapers.

International web-based newspapers have been analyzed by Scharl and Bauer (2004), specifically they identified key terms related to solar power technologies. They could make a contingency table of the number of occurrences of these terms in the newspapers. And using this table they used correspondence analysis (CA) to jointly plot the terms and newspapers. This provided an overview of the relationships of terms and newspapers related to solar power technologies. In addition they provided the movement of the newspapers over time by applying CA on multiple periods using the same terms.

This thesis also focuses on CA, similar to what Scharl and Bauer did. That is, by analyzing the occurrences of political parties in newspapers. The CA solutions over time are examined as well, but applied on the co-occurrences of terms with political parties.

In the paper by Šilić et al. (2012) a method is described based on CA called CatViz. First, they found terms written in newspapers over a period of time and applied CA. Then, they jointly plotted the terms and years, focusing specifically on the names of persons and entities. The reason for this is that this carries most of the information on the who and where questions. The goal of the method was to obtain a knowledge discovery tool. To proof its usefulness, they did an experiment with 11 persons to find out whether they could gain knowledge. It turned out that after using it these persons were indeed more knowledgeable about the world events of the last 20 years.

As opposed to this paper, in this thesis I leave out the names of persons. I do take entities into account, specifically the political parties. The focus is on events and as many of the persons are affiliated with political parties, this distorts the analysis.

Shineha et al. (2008) have analyzed Japanese newspapers using multiple CA to detect topics on genetic modification covered in newspapers over the years. This way they found certain shifts on the topic. Namely, the change of genetic modification's application and the public opinion on the topic.

Both Shineha et al. and Šilić et al. use CA on the occurrences of terms within subsequent time periods. As a knowledge discovery tool I also use this type of analysis to explore the use of terms for each period.

Sometimes the choice of terms is arbitrary. Zuell (2008) describes a technique to identify events as a list of terms, she calls this the reference text technique. She qualified a major event

as one where the following must apply: "The news are reported for at least several days in mass media and give rise to widespread public discussion and/or to a substantive increase in media use. Additionally, major events attract wide and long-lasting attention." After lemmatization (finding the stem and synonyms of each word) she looked at the relative term frequencies of the whole dataset of terms and of a smaller time period. Then she looked at the difference between these relative frequencies and selected those with the highest relative differences. This way she was able to discover specific events. Subsequently she applied exploratory factor analysis and multiple CA which uncovered the major events that happened in the smaller time period. Because of its simplicity and effectiveness, I find the relevant terms for analysis by applying this technique in my research.

Using CA, I discover relationships and associations between terms, political parties and newspapers. To summarize, I apply CA in three different ways. The first one as described by Scharl and Bauer (2004), by making a frequency table of the number of occurrences of terms within each newspaper. The second way is similar to what Šilić et al. (2012) and Shineha et al. (2008) have done with the frequencies of terms in each time period. Finally, I examine the co-occurrences of political parties with the relevant terms. I believe this to be a novel application, as I have not found this way of using CA on term frequencies in the literature.

3 Methodology

In this chapter I explain the methods used in my analysis. In section 3.1, I explain the data collection method. Here I also show how to prepare these for analysis. I extract relevant terms from the texts within a time period using the reference text technique (Zuell, 2008), which is explained in section 3.2. I explain the use of correspondence analysis (CA) in section 3.3, in addition I show several ways to interpret the results of this method. To rotate a CA configuration as close as possible towards another CA configuration, I explain the Procrustes rotation method in section 3.4.

3.1 Preparing the data

For this research I retrieved articles from several major Dutch national newspapers as listed in Table 1. I obtained articles from these newspapers using the online LexisNexis Academic resource. For this purpose, I selected all articles containing one or more of the names of all the political parties which are represented in the Dutch parliament, also listed in Table 1. The retrieved articles always mention at least one political party, the resulting list of articles consequently covers mostly national political news.

Table 1: The political parties and their number of seats in the Dutch Parliament after the 2012 elections (source: Tweede Kamer der Staten-Generaal). And the newspapers with their circulation numbers over 2012 (source: HOI, Instituut voor Media Auditing)

Parties	Seats
Volkspartij voor Vrijheid en Democratie - VVD	41
Partij van de Arbeid - PvdA	38
Partij voor de Vrijheid - PVV	15
Socialistische Partij - SP	15
Christen Democratisch Appel - CDA	13
Democraten 66 - D66	12
Christenunie - CU	5
GroenLinks - GL	4
Staatkundig Gereformeerde Partij - SGP	3
Partij voor de Dieren - PvdD	2
50PLUS	2

Newspapers	Circulation
De Telegraaf	582,582
Metro	439,402
Algemeen Dagblad (AD)	420,977
Sp!ts	323,974
De Volkskrant	260,708
NRC Handelsblad	199,359
Trouw	104,155
NRC.NEXT	79,387
Het Financieele Dagblad (FD)	54,678
Reformatorisch Dagblad (RD)	50,248
Nederlands Dagblad (ND)	26,039

The obtained articles from LexisNexis arrives as a list. In order to analyze it, this has to be transformed to a table. Therefore, I used and adjusted a script written by Caren (2012) in the Python programming language. This script extracts each article by obtaining its title, the contents, the date and other properties. See appendix C for the code of the script.

I retrieved articles from the period July 2010 until June 2013. The resulting dataset contains articles as rows. The columns represent the variables, including the article text, published date, originating newspaper, type of article and number of words of the article. Moreover, from the dates, I created variables for the months and quarters.

Before extracting the terms from each article, a number of transformations have to be done on the texts. First, punctuation has to be removed. Second, all letters have to be turned lowercase. Third, stopwords have to be deleted. These stopwords are the most frequently used words in Dutch (e.g. 'van', 'de', 'het', 'in'). For the complete list of these words, see appendix A. Finally, full names of political parties are abbreviated. This provides consistency in the names. So 'partij van de arbeid' becomes 'pvda', 'partij voor de dieren' becomes 'pvdd', and 'partij voor de vrijheid' becomes 'pvv', but also, 'cu' becomes 'christenunie'.

To extract the terms used in each article, I used a so-called word tokenizer. This extracts terms which are between spaces and/or ends with a question mark, exclamation mark, punctuation mark or semicolon. These terms can be represented in an $n \times m$ matrix X with the selected m terms in the columns and the n newspaper articles as the rows. Each element X_{ij} represents the number of occurrences of term j in article i. This is called a document-term matrix. An example is shown in Table 2.

Table 2: Example of a document-term matrix

Article	cda	d66	pvda	pvv	vvd	illegaal	langstudeerboete	lenteakkoord
1	1	0	0	0	0	2	0	1
2	0	0	3	1	2	6	0	0
3	2	3	0	0	1	0	3	2

3.2 Extracting relevant terms

Using the document-term matrix, I determine the relevant terms for a certain time period, by applying the reference text technique described by Zuell (2008). The idea is that a term that occurs more often in a certain time period as opposed to a larger reference time period is regarded as being related to a key event.

The dataset of terms for the larger reference time period is included in the $n \times m$ document-term matrix X. Furthermore, X_t is a subset of X denoting period $t = 1, ..., P$, with P the total number of periods. Then one looks at the term frequencies of X and X_t as a part of the sum of its totals, that is: $\sum_i X_{ij,t} / \sum_{i,j} X_{ij,t}$ and $\sum_i X_{ij} / \sum_{i,j} X_{ij}$. The differences between these two terms are:

$$d_{j,t} = \frac{\sum_i X_{ij,t}}{\sum_{i,j} X_{ij,t}} - \frac{\sum_i X_{ij}}{\sum_{i,j} X_{ij}} \tag{1}$$

For each term j and period t, a value $l_{j,t}$ is obtained using the following formula:

$$l_{j,t} = \frac{d_{j,t}}{\sum_{i,j} X_{ij,t}} \tag{2}$$

Using this, the highest values of $l_{j,t}$ correspond to the terms which signify events in that period. These will be used in a new document-term matrix.

To limit the number of terms, I remove the terms occurring zero times in 99% of the documents. This percentage is arbitrary, but should be high enough to obtain enough terms, and low enough to not include irrelevant terms.

In my research, these events are in a large part related to national political news, because of the nature of the dataset. Therefore, the selected terms include political matters that were of particular interest in a certain time period.

3.3 Correspondence Analysis

Correspondence analysis (CA) has been developed in France (Benzécri, 1973) and been made popular by, amongst others, Greenacre (1984). Given a $p \times q$ frequency matrix N, CA allows the visualization of this data in a two or sometimes three dimensional plot. In this plot, the rows and columns of the frequency matrix are plotted as points. This provides a way to efficiently interpret the frequency matrix, making it easier to find associations and relationships between the rows and columns.

In this section, I first describe the correspondence matrix in section 3.3.1, after which I introduce the singular value decomposition (SVD) in section 3.3.2, which is an important part of the calculations for CA. After that, I explain the calculations of the CA coordinates in section 3.3.3 and how to interpret CA in section 3.3.4. Then I go into detail on how to add supplementary points in section 3.3.6 and how to diagnose points using contributions in section 3.3.5. To measure the stability of a CA solution I use bootstrap methods, explained in section 3.3.7. Finally, in section 3.3.8 I outline the frequency matrices on which I have used CA to analyze them.

3.3.1 Correspondence Matrix

The $p \times q$ correspondence matrix P is obtained by dividing each element of the frequency matrix N by the grand total of N. The row and column masses r_i and c_j are calculated as the sum of the elements of row i and column j:

$$r_i = \sum_j p_{ij}, \text{ where } i = 1, ..., I \tag{3}$$

$$c_j = \sum_i p_{ij}, \text{ where } j = 1, ..., J \tag{4}$$

or in matrix notation:

$$r = P1_q, \; c = P'1_p \tag{5}$$

where 1_x denotes a vector of ones with length x.

Each row of P divided by its sum is called a row profile. This describes the row in terms of the relative frequencies of its total per column. Similarly, each column of P divided by its sum is called a column profile.

3.3.2 Singular Value Decomposition

Given any data matrix, using the singular value decomposition (SVD) property there are two important applications. First, it transforms correlated variables in this matrix to new uncorrelated variables. Second, these new variables are found and ordered in such a way to have most of the variance across each dimension. This can reduce the dimensionality of the data, thereby simplifying the analysis of the data.

Given any matrix S, SVD decomposes it into two orthogonal matrices U and V and a diagonal matrix Λ containing singular values λ_k in descending order:

$$S = U\Lambda V', \tag{6}$$

with $U'U = V'V = I_k$, Λ diagonal and $\text{rank}(S) = k$ Geometrically, it scales the matrix U with Λ and applies the counter-clockwise rotation V'. Turning this around and rewriting the expression in terms of U we get

$$U = SV\Lambda^{-1} \tag{7}$$

So, S is applied an orthogonal rotation by V and is scaled by Λ^{-1}. The rotation and scaling aligns and normalizes the data points along each dimension. This is done in such a way that most of the variance is found and ordered over each dimension.

A useful application of the SVD is the Eckart-Young Theoreom (Eckart and Young, 1936). This states that one can construct a matrix $S_{(m)}$ out of the first m columns of $U_{(m)}$ and $V_{(m)}$ with the first m singular values in $\Lambda_{(m)}$. That is, $S_{(m)} = U_{(m)}\Lambda_{(m)}V'_{(m)}$. Then $S_{(m)}$ is a least-squares approximation of S.

3.3.3 Calculation of Coordinates

Given the correspondence matrix P, the matrix of expected values of P is calculated using the probabilities of seeing the corresponding row with the corresponding column, which are contained in r and c. So the expected values are found by rc'. By finding the differences between each element of P and rc', we find the residuals of this estimate. Dividing these residuals by the square root of rc', we obtain the standardized residuals:

$$s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j} \tag{8}$$

or in matrix form:

$$\tilde{P} = D_r^{-\frac{1}{2}}(P - rc')D_c^{-\frac{1}{2}} \tag{9}$$

with $D_r = diag(r)$ and $D_c = diag(c)$ diagonal matrices with elements of respectively r and c on its diagonals. We want to minimize the standardized residuals. So CA aims to find coordinate matrices F for the row and G for the column points in such a way that the loss function

$$\phi(F, G) = \|\tilde{P} - D_r^{\frac{1}{2}} F G' D_c^{\frac{1}{2}}\|^2 \tag{10}$$

is minimized (van de Velden and Kiers, 2005). Here, $\|A\|^2$ is defined as the sum of the squared elements of A.

The singular value decomposition of S is

$$S = U\Lambda V' \tag{11}$$

with Λ containing the singular values λ_k in descending order. Because of the theorem by Eckart and Young (1936) as described in the previous section, the first k dimensions approximate S. Using this, the so-called standard coordinates for the rows and columns can be obtained, they are respectively

$$\phi_{ik} = u_{ik}/\sqrt{r_i} \tag{12}$$

and

$$\gamma_{jk} = v_{jk}/\sqrt{c_j}. \tag{13}$$

The principal coordinates scale the standard coordinates by the singular values λ_k. So that for the rows and columns these coordinates are given by

$$f_{ik} = \lambda_k \phi_{ik} \tag{14}$$

and

$$g_{jk} = \lambda_k \gamma_{jk}. \tag{15}$$

A third way is provided by scaling the row and column coordinates similarly. This consitutes a symmetrical biplot, the interpretation of which is explained later. The coordinates are given by

$$\hat{f}_{ik} = \sqrt{\lambda_k}\phi_{ik} \tag{16}$$

and

$$\hat{g}_{jk} = \sqrt{\lambda_k}\gamma_{jk}. \tag{17}$$

In matrix notation, the coordinates can be written as:

$$F = D_r^{-\frac{1}{2}} U_k \Lambda_k^{\alpha} \tag{18}$$

$$G = D_c^{-\frac{1}{2}} V_k \Lambda_k^{1-\alpha} \tag{19}$$

Then with different choices of α, the coordinates matrices F and G can be plot with different

interpretations (van de Velden and Kiers, 2005). With $\alpha = 0$ (or $\alpha = 1$), the plot is called an asymmetric plot, with the column (row) coordinates as principal coordinates and the row (column) coordinates as standard coordinates. This case is also called column-principal (row-principal) normalisation. When $\alpha = \frac{1}{2}$, the plot is referred to as a symmetric CA biplot. The row and column coordinates are both scaled equally. This setting is also called symmetrical normalisation. Finally, both the rows and columns can be plotted using only the prinicpal coordinates. This is called a symmetric CA plot or CA with principal normalisation.

3.3.4 Interpretations

For readibility, I will explain the interpretations of CA by looking at the row profiles, however the same applies to the column profiles. In the CA solution, each of the first dimensions contains the maximized explained differences. Therefore, each dimension has their own interpretation, uncorrelated with other dimensions. This helps to better understand the relationships between the rows and columns and between each point.

Plotting the asymmetric plot with row-principal normalisation, relative Euclidean distances between the row points are interpreted as the amount of dissimilarity between the corresponding row profiles. That is, a smaller distance means a larger similarity and a larger distance a smaller similarity. This only applies between the row points and not when comparing row points with column points.

The asymmetric plot, but also the symmetrical biplot, constitute a biplot. In a biplot, the points of the column and row profiles are represented as scalar products $x'y$. The angle between the column and row vector is equal to θ. Following geometry calculations the following applies:

$$x'y = x \cdot y \cdot cos(\theta) \tag{20}$$

This means that one can project the column point onto the vector of a row point. Then the value of this projection is multiplied by the length of the row vector, which is equal to the scalar product. These are approximations of the standardized residuals (Greenacre, 2010).

The total inertia of a CA solution is found as the χ^2-statistic of the frequency table divided by the total sum of the table: $\frac{\chi^2}{n}$. So it is dependent on this χ^2-statistic, which is a measure of the similarity between row or column profiles. Therefore, with a higher inertia, there are more differences between the profiles, which leads to the points being more spread out in the CA plot. With each kth dimension, part of the inertia is explained as each kth principal inertia. This is calculated as the square of the singular values: λ^2 and is represented as a percentage of the total inertia. It shows how much of the frequency table is explained in the CA solution per dimension.

3.3.5 Contributions

Within CA, each principal axis makes a contribution to the total inertia, called the principal inertia. It is calculated as the sum of the mass of the ith row r_i times the squares of the

principal coordinates f_{ik}: $\sum_i r_i f_{ik}^2$. Similarly, each row makes a contribution to the total inertia by $r_i \sum_k f_{ik}^2$. For the kth axis, this contribution is $r_i f_{ik}^2$. A division is made between the absolute and relative contributions. The absolute contribution is obtained by expressing $r_i f_{ik}^2$ relative to the principal inertia:

$$\omega_{ij} = \frac{r_i f_{ik}^2}{\sum_i r_i f_{ik}^2} = \frac{r_i}{\lambda_k^2} f_{ik}^2 \tag{21}$$

It is a measure for how each point contributes to the principal axis. Therefore, knowing which points have a high absolute contribution helps the interpretation of that axis. The relative contribution is obtained by expressing $r_i f_{ik}^2$ relative to the row inertia:

$$\sigma_{ij} = \frac{r_i f_{ik}^2}{r_i \sum_k f_{ik}^2} = \frac{f_{ik}^2}{\sum_k f_{ik}^2} \tag{22}$$

This is a measure for how well the point is represented on the principal axis (Greenacre, 1993).

3.3.6 Supplementary points

Each row and column profile contributes to the CA configuration. However, it is also possible to include supplementary points, which are projections onto the CA plot, but have no influence on the CA solution itself. The position of these points are determined similar to the original points. When a supplementary row with values in g is added, first each element g_j is divided by its total $\sum_l g_l$ to get $g_j / \sum_l g_l$. Then the supplementary row coordinates $\hat{\phi}_{ik}$ are found by multiplying this with the standard coordinates of the columns:

$$\hat{\phi}_{ik} = \sum_l \frac{g_j}{g_l} \gamma_{jk} \tag{23}$$

Supplementary points do not contribute to the CA solution, so also not to the position of the axes. Therefore, they are not provided with a value for the absolute contribution.

3.3.7 Bootstrapping

Within a CA solution, the points are given a definite place in the plot. However, this doesn't give information about the stability of that point in the plot. In other words, if something changes in the data, how does this affect each individual point? Also, what is the true position of the point? Using bootstrapping, a measure is obtained for the quality of the CA in terms of stability and bias (van de Velden et al., 2012).

A random sample is drawn B times from the document-term matrix X with replacement and with the same length as X, this is also called the bootstrap sample. In one sample, items can therefore be drawn more than once or not at all. For each of the B bootstrap samples, CA is applied and the solution saved. The CA solution finds the coordinates with the axes explaining the largest part of the inertia. However, these solutions are not always similar.

As per Equations 9 and 10, the CA solution $\hat{P}$ is $\hat{P} = D_r^{\frac{1}{2}} F G' D_c^{\frac{1}{2}}$. By rotating both F and G with rotation matrix T, this results in

$$\hat{P} = D_r^{\frac{1}{2}} F T T' G' D_c{}^{\frac{1}{2}}. \tag{24}$$

As a rotation matrix, T is orthogonal and $T'T = TT' = I$, so that Equation 24 again becomes

$$\hat{P} = D_r^{\frac{1}{2}} F G' D_c{}^{\frac{1}{2}}. \tag{25}$$

Therefore, any possible rotation is allowed and the CA solutions of the bootstrap samples have to be rotated towards one single solution. This can be done by applying Procrustes rotation (described in section 3.4) to rotate the points towards the original CA coordinates. From these final bootstrapped CA solutions, the principal row and column coordinates f and g are obtained.

The rotation is applied towards the original CA coordinates, which are not necessarily the 'real' population coordinates. Therefore, this might lead to the bootstrap coordinates to be closer to these original coordinates and also to each other. Consequently, there might be some underestimation of the variance in the bootstrap points (Ringrose, 1996). So the bootstrapped sample doesn't include the real movements of points within the dimensions. In this case, it only visualizes the stability of the relative differences between the points.

The bootstrapped row and column points can be plotted to visualize this stability. However, with a large amount of points, this would make it less readible. It is possible to calculate 90% confidence ellipses and to draw these around the points. This makes it easier to judge the stability directly from the plot.

The bias of each point is found as the difference between the original CA coordinates and the centroid of the bootstrapped points. This bias can be made visible by drawing a line from this centroid to the original point.

With a large amount of points, the confidence ellipses can make the plot unreadable. In order to assess the stability of each point, the mean squared error (MSE) can be calculated (van de Velden et al., 2012). It accounts for both the variance and bias of a point and is found as the mean of the squared difference between rotated bootstrap coordinates and the original CA coordinates. For the rows we get

$$MSE_i = \frac{1}{B} \sum_{b=1}^{B} (f_{ib} - \hat{f}_i)'(f_{ib} - \hat{f}_i) \tag{26}$$

where f_{ib} is the principal row coordinate of the CA on the bth bootstrap sample, $\hat{f}_i$ is the principal row coordinate of the original CA solution and B is the number of bootstrap samples.

Similarly, for the columns:

$$MSE_j = \frac{1}{B}\sum_{b=1}^{B}(g_{jb} - \hat{g}_j)'(g_{jb} - \hat{g}_j) \tag{27}$$

where g_{jb} is the principal column coordinate of the CA on the bth bootstrap sample and $\hat{g_j}$ is the principal row coordinate of the original CA solution.

To compare the stability measures between different solutions, MSE is not sufficient. One can use the relative mean squared errors (RMSE) instead. To arrive at this, first introduce the total sum of squares (TSS) as

$$TSS_{rows} = \sum_{i=1}^{n} \hat{f}_i'\hat{f}_i \text{ and } TSS_{columns} = \sum_{j=1}^{n} \hat{g}_j'\hat{g}_j \tag{28}$$

Then for both the rows and columns, the RMSE is defined as the sum of the MSEs divided by the TSS to get

$$RMSE_{rows} = \frac{\sum_{i=1}^{n} MSE_i}{TSS_{rows}} \tag{29}$$

and

$$RMSE_{columns} = \frac{\sum_{i=1}^{n} MSE_j}{TSS_{columns}}. \tag{30}$$

A measure for the overall RMSE of the solution can be found by finding the average of the RMSEs for the rows and columns:

$$RMSE = \frac{1}{2}(RMSE_{rows} + RMSE_{columns}) \tag{31}$$

3.3.8 Frequency Matrices

I applied CA on three different frequency matrices: terms mentioned per time period, terms mentioned per newspaper and parties mentioned with each term. Moreover, for the latter I also consider different frequency matrices over time per quarter.

The document-term matrix obtained in section 3.1 contains the frequencies of terms in each document. Knowing which newspapers and time periods these documents occur in, it is easy to get the frequencies of terms (including parties) per newspaper and per time period from this matrix by adding the counts together.

To arrive at a frequency table describing co-occurrence of political parties with other terms, I apply a simple matrix multiplication in the following way. Given the $n \times m$ document-term matrix X, excluding parties within the terms, and $\tilde{X}$ an $n \times p$ document-term matrix with only the p parties as terms and in binary format, occurring or not occurring, then $X'\tilde{X}$ gives an $m \times p$ matrix with the number of times a term is mentioned with a party. An example of such a frequency matrix is shown in Table 3.

Table 3: Example of a frequency matrix

	cda	d66	groenlinks	pvda	pvv	vvd
langstudeerboete	1	6	9	2	4	0
illegaal	3	1	0	4	3	0
lenteakkoord	9	8	2	1	3	7

3.4 Procrustes Rotation

As a result of the Eckart-Young theorem (Eckart and Young, 1936), CA maximizes the inertia for each subsequent dimension. Therefore, the row and column profiles are positioned in such a way that it is possible to interpret each axes separately. Viewed in this way, each point's 'definite' position is of interest. When obtaining CA solutions for different time periods, each CA plot provides different definite positions for each point. This makes it difficult to compare the plots of each period. We are no longer interested in the definite positions as such, but will examine the 'relative' positions. This is achieved by rotating each CA solution in such a way that its points' coordinates are as close as possible to the coordinates of the previous period's CA solution.

This is possible using Procrustes rotation (Hurley and Cattell, 1962), which rotates and mirrors a $u \times v$ coordinates matrix Y so that its points will be closer to or on the same place as a target $u \times v$ matrix Y^*. Procrustes rotation can also incorporate scaling and translation of the configuration. By scaling Y for each period, the scales of each period's solutions are not similar. By translating, the positions of each point relative to the axes changes for each period. Therefore, both scaling and translation affect the comparison between the CA solutions of each period. Consequently, I will not include these and only describe the orthogonal Procrustes.

3.4.1 Procedure

When rotating, Y is multiplied with the $u \times v$ rotation matrix T, so that $YT \approx Y^*$. The equation that provides a rotation T to Y to arrive at the new rotated matrix $\tilde{Y}$ is then as follows:

$$\tilde{Y} = YT \tag{32}$$

where 1 is an $n \times 1$ vector of ones. The aim is to let $\tilde{Y}$ be approximately the same as the original matrix Y^*: $\tilde{Y} = YT \approx Y^*$. To arrive at this, minimize $Y^* - YT$, which is done by minimizing the sum of squared differences between the two terms,

$$L(T) = \text{tr}(Y^* - YT)'(Y^* - YT). \tag{33}$$

Expanding this,

$$\begin{aligned} L(T) &= \operatorname{tr}(Y^* - YT)'(Y^* - YT) \\ &= \operatorname{tr} Y^{*\prime}Y^* + \operatorname{tr} T'Y'YT - 2\operatorname{tr} Y^{*\prime}YT \\ &= \operatorname{tr} Y^{*\prime}Y^* + \operatorname{tr} Y'Y - 2\operatorname{tr} Y^{*\prime}YT \end{aligned}$$

where $T'T = TT' = I$. Since $L(T)$ is minimized over T, and the first two terms are not dependent on T, this equation is simplified to

$$L(T) = c - 2\operatorname{tr} Y^{*\prime}YT \tag{34}$$

where c is a constant not dependent on T.

The singular value decomposition of $Y^{*\prime}Y$ is found, which is $U\Lambda V'$. Here, $U'U = V'V = I$ and Λ is the diagonal matrix with singular values. Using this, Equation 34 becomes

$$\begin{aligned} L(T) &= c - 2\operatorname{tr} Y^{*\prime}YT = c - 2\operatorname{tr} U\Lambda V'T \\ &= c - 2\operatorname{tr} V'TU\Lambda \\ &\geq c - 2\operatorname{tr}\Lambda. \end{aligned}$$

This last inequality is similar to a lower bound inequality derived by Kristof (1970):

$$-\operatorname{tr} ZW \geq -trW \tag{35}$$

with equality if and only if $Z = I$. Then $Z = V'TU$ and $W = \Lambda$ and $L(T)$ is minimal when $Z = I$ or $V'TU = I$. This holds when

$$T = VU', \tag{36}$$

which is then the Procrustes rotation matrix T (Borg and Groenen, 2005).

3.4.2 CA Contributions After Rotation

After rotating a CA solution, the axes are rotated with it and thereby its interpretations using the principal inertias and the contributions. Here I give the explanation for obtaining these contributions once again (van de Velden and Kiers, 2005).

The principal inertia of the rotated solution can be calculated as

$$\tilde{\Lambda}^2 = (T_c\Lambda^2 T_c') \odot I_K \tag{37}$$

Here, T_c is defined as a $K \times K$ matrix with the rotation matrix T in the upperleft corner and an identity matrix in the rest of the downright corner and $\odot$ denotes the elementwise product.

The absolute and relative contributions for the rows (as principal coordinates) are given by

$$\tilde{\omega}_{ij} = \frac{r_i}{\tilde{\lambda}_k^2} \tilde{f}_{ik}^2 \tag{38}$$

and

$$\tilde{\sigma}_{ij} = \frac{\tilde{f}_{ik}^2}{\sum_k \tilde{f}_{ik}^2}. \tag{39}$$

For the columns, which are the standard coordinates in this case, it is not possible to obtain the relative contributions in a meaninful way. This is because as the axes rotated, the basis for the calculations of the coordinates is no longer orthogonal. The absolute contributions, however, are obtainable, similar to equation (38), but by considering each axis separately, that is, as if it is the only axis.

3.4.3 Congruence Coefficient

When applying the Procrustes rotation on a CA solution, one wants to examine whether each rotated solution truly looks somewhat similar to its previous CA solution. If this is not the case, and the differences are too large, the relative points of each solution will be located in different locations. This makes comparison between the time periods more difficult. For this comparison, Tucker's congruence coefficient is helpful as a measure for the similarity in distances between two configurations (Abdi, 2007).

First, the Euclidean distances between each point's coordinates is the Pythagorean distance between two points u_i and u_j. It is calculated by finding the square-root of the sum of squared differences between each coordinate:

$$D_{ij}(U) = \sqrt{\sum_a (u_{ia} - u_{ja})^2} \tag{40}$$

Given the Euclidean distance matrix D, and similarly for another distance matrix D^*, the congruence coefficient between these two configurations is given by

$$\phi = \frac{\sum_{ij} D_{ij} D_{ij}^*}{\sqrt{\sum_{ij} D_{ij}^2 \sum_{ij} D_{ij}^{*2}}} \tag{41}$$

A rule of thumb is that if the congruence coefficient of two solutions is between than 0.85 and 0.95, they are considered as having fair similarity (Lorenzo-Seva and Ten Berge, 2006). In this case, if this number is too high, it would mean that there is not much movement between each period. Therefore, a congruence coefficient which is a bit lower than 0.85 is still an acceptable value.

4 Results

In this chapter I discuss the obtained results. First I introduce the retrieved data in section 4.1. Then in section 4.2, I explain the extraction of relevant terms from the data. Finally in section 4.3 I show and give interpretations to the different obtained CA results.

4.1 Data

The complete dataset that I downloaded from LexisNexis consists of 114.785 newspaper articles from the period July 2009 until June 2013. I apply the correspondence analyses using the data from the period July 2012 until June 2013. This consists of 29.253 articles.

Since I deal with articles found using the names of political parties, these articles are mostly related to political news. The total number of these articles per newspaper are shown in the graph on Figure 1. Except for the 'Metro' and 'Spits', the newspapers with higher circulation numbers (see Table 1) also have a larger number of political articles.

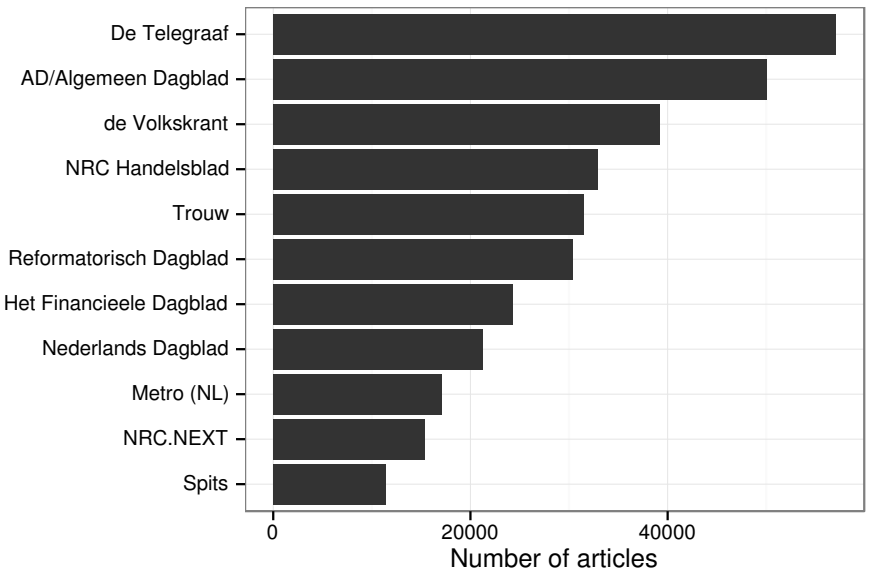

Figure 1: Total number of articles about Dutch politics per newspaper over the period July 2012 - June 2013

The graph in Figure 2 shows the number of articles about politics per newspaper as a percentage of its total in the period July 2012 until June 2013. Especially 'NRC Handelsblad' has many articles on politics and the Christian newspapers 'Nederlands Dagblad', 'Reformatorisch Dagblad' and 'Trouw' write much about it. 'De Telegraaf', 'Metro' and 'Algemeen Dagblad' are the newspapers that write relatively little about the political parties, even though they have a large amount of articles.

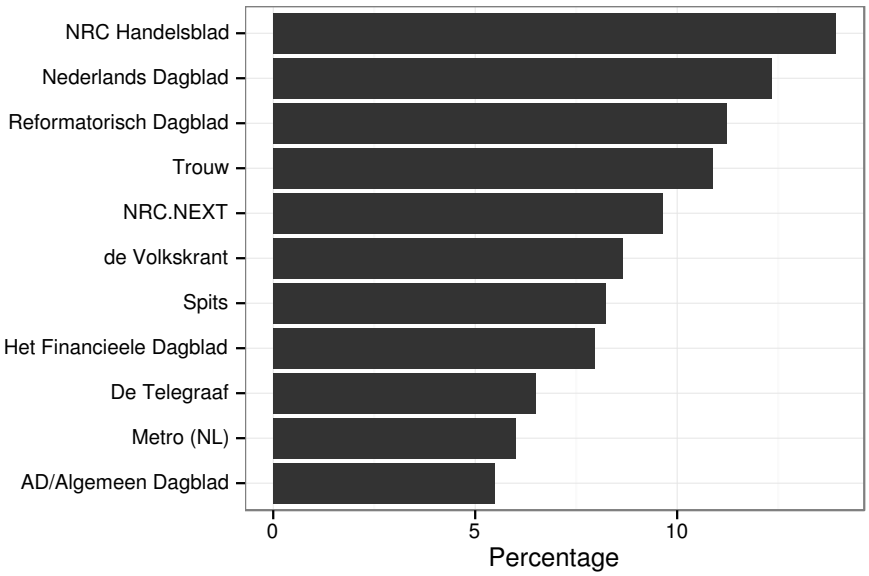

Figure 2: Number of articles about Dutch politics as a percentage of the total number of articles over the period July 2012 - June 2013

4.2 Extracting relevant terms

Following the procedure explained in section 3.2, I extract the 100 terms with the highest values of $l_{j,t}$ (see Equation 2. I did this for the period July 2012 until June 2013 and separately for each quarter in this period. What results is a list of terms associated with events for each of those periods. These resulting sets of terms contain items that have too much assocation with specific parties or are not relevant for the analysis and I therefore removed these. They included:

- Names of political party members (e.g. 'Mark Rutte', 'Lodewijk Asscher', 'Geert Wilders')
- Terms that were commonly mentioned with each party (e.g. 'verkiezingsprogramma', 'lijsttrekker', 'partij')
- Names of places in the Netherlands (e.g. 'Roermond', 'Goeree-Overflakkee', 'Lansingerland')

In addition, I replaced synonyms and similar terms (e.g. 'troonswisseling' and 'abdicatie', 'inkomensafhankelijke' and 'inkomensafhankelijk'). Finally, I constructed a new document-term matrix including only the 30 terms with the highest values of $l_{j,t}$. For the full list of these, see appendix B.

4.3 Correspondence Analysis

In this section I discuss the obtained CA results. I plotted each CA solution as two-dimensional symmetric and asymmetric plots. The resulting plots bring insight into the relations between newspapers, terms and political parties. In these plots, for each row and column profile, the absolute contributions are shown as large or small. A large contribution in this case means that it is larger than the average absolute contributions over that dimension. These points contribute relatively more to that axis than other points, which therefore show the importance of those points within that axis.

The CA plots I show, have different settings and thus different interpretations. I use iconography to make it easier to understand the interpretations for each plot (Gower et al., 2014). The icon for the plot having shape parameter 1 makes clear that units along both axes are the same. When an icon has two similar (or different) points connected with a line, it means that the distance between the same (different) points are interpretable. If the icon includes an angle connected with two points, it means that the projections of one point to the second (same or different) are interpretable. A slash through the icon means that the corresponding interpretation is not valid.

This section is structured in the following way. First of all, in section 4.3.1, I discuss the CA with the newspapers on the columns and the parties mentioned in each newspaper on the rows. In section 4.3.2, I discuss the CA with the terms on the rows, mentioned in each time period on the columns. In section 4.3.3 and 4.3.4 I will go into the results for the co-occurrences of terms (rows) with parties (columns) in the whole period July 2012 until June 2013 and of each quarter in this period, respectively.

4.3.1 Parties by newspapers

The CA plot with the Dutch newspapers by the political parties are shown in Figure 3. It includes the confidence ellipses for both the political parties (columns) and the newspapers (rows). The CA solution is found using $\alpha = \frac{1}{2}$ in Equation 18. This makes it possible to interpret the relations between the rows and columns using a biplot, while still keeping the plot readable.

The plot shows a division between the parties and newspapers mostly into coalition, opposition and Christian parties. The 'NRC Handelsblad' and 'Telegraaf' write mostly about the coalition parties than other newspapers. The Christian parties are more written about by 'Nederlands Dagblad', 'Reformatorisch Dagblad' and 'Trouw' and the other newspapers divided their attention on the other opposition parties.

Originally, I also included the two newspapers 'Nederlands Dagblad' and 'Reformatorisch Dagblad'. However, both had a large absolute contribution on the first dimension and were located on the far edge of the plot, while also having a relatively small frequency. They caused the other points to be mostly clustered together and didn't give much extra information. Therefore, I considered them as outliers and left these two newspapers out of the analysis as supplementary points.

The resulting CA has a cumulative explained inertia of 80.4%. This high percentage means that the plot explains most of the variance quite well. The first horizontal dimension shows a difference between the coalition parties (VVD and PvdA) and opposition parties (D66, PVV, SP, Groenlinks, ChristenUnie, SGP, PvdD and 50PLUS). The CDA was a ruling party in the previous government, which might explain it being located on the coalition side. The second dimension can be divided into Christian parties (CDA, ChristenUnie and SGP) on one side and the Socialist Party (SP) on the other. Most parties have little to no overlap with each other, except for the SGP and ChristenUnie. Both of these parties are conservative Christian parties. This also confirms the interpretation of the top-part focusing on Christian parties.

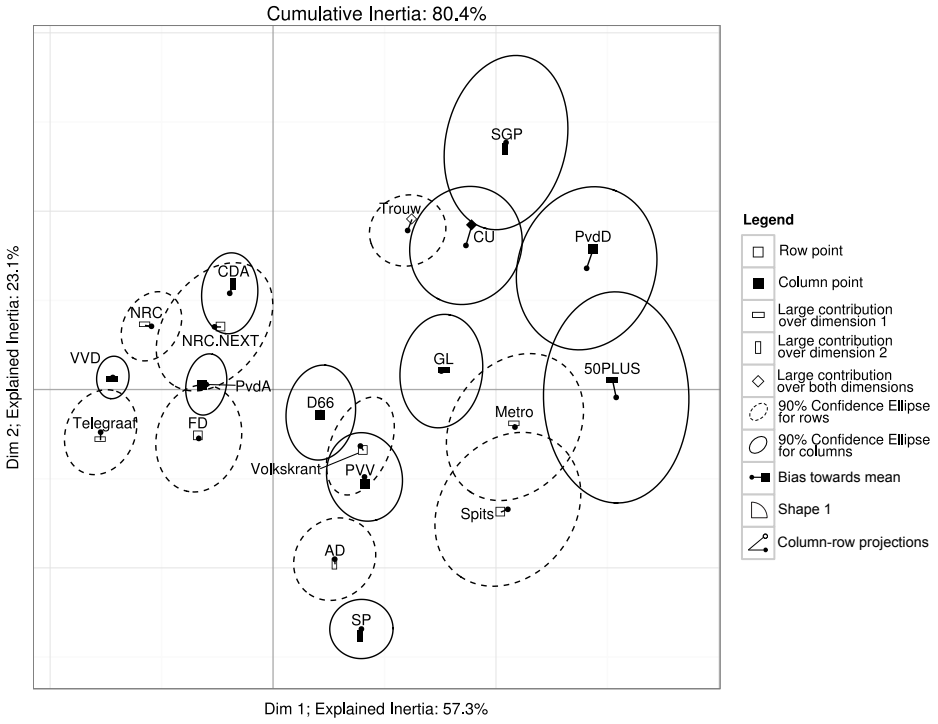

Figure 3: Correspondence analysis solution with symmetric normalisation. The solution is calculated from the frequency table of the occurrences of political parties (columns) in Dutch newspapers (rows). The confidence ellipses are calculated using a bootstrap on the CA solution. For this solution, the RMSE is 0.4746.

Given these interpretations, we can observe how newspapers relate to these axes and to each other. We see that 'NRC Handelsblad' and 'De Telegraaf' write relatively more about the coalition, especially the VVD. Also 'Het Financieel Dagblad' and 'NRC.NEXT' cover these parties relatively much. There is some overlap in the confidence ellipses of 'NRC.NEXT' with 'NRC Handelsblad'. This can easily be explained by the overlap of articles in both papers, since they are made by the same publisher. Also, there is a little overlap with 'Financieel Dagblad'.

On the opposite side, 50PLUS, GroenLinks and ChristenUnie have the highest contributions. These are most of the smaller parties in the Parliament. Especially 'Metro' and 'Trouw' write about these opposition parties. But also 'Spits' seem to write relatively more about them. A large overlap in confidence ellipses can be found for 'Spits' and 'Metro'. This is most likely because the articles in these two newspapers are mostly from the same news agency (ANP).

On the Christian upper part of the plot, the only newspaper clearly present is 'Trouw'. 'Nederlands Dagblad' and 'Reformatisch Dagblad' are located in the most upper part of the plot as supplementary points, with 'Reformatisch Dagblad' placed on top (these are not included because they extend the plot too much). What is interesting to note is that the PvdD is also located in this upper part. The lower part is dominated only by the SP. The single newspaper that writes relatively more about this party seems to be 'Algemeen Dagblad'. 'Spits' also tends somewhat towards this part of the plot. Finally, 'de Volkskrant' is placed somewhat in the lower-left part of the plot, tending towards opposition parties, but not the Christian ones. It seems to be focused more on the PVV and D66.

4.3.2 Terms by time period

Figure 4 shows the CA of the quarters by the terms, including political parties. From Equation 18, $\alpha = 1$, so this is an asymmetric plot with row-principal normalisation. This way the terms as the row points can be compared to each other. In addition, the row points can be related to the time period using the biplot. In the plot, the bias towards the mean of the confidence ellipses is not shown. It is close to zero for all terms and periods and would make the plot less readable.

We can see that three time periods can be distinguished. The third quarter of 2012 can be found in the lower half, which indicates the election period. The fourth quarter of 2012 is located in the upper right corner, which is the beginning of the ruling period. Finally, the first half year of 2013 is in the upper left corner, the second part of the ruling period. The coalition parties are shown to be relatively more in the news during the beginning of the ruling period. The opposition parties have their focus on the election period, where most got more attention in the second part of the ruling period.

Observing the position of the political parties, the VVD and PvdA are clearly located in the upper right part. The D66, ChristenUnie and SGP are close to each other on the left side, and 50PLUS is in that same direction but farther away. The other parties are placed in the lower part. This means that during the election period these parties were mentioned relatively more often as compared to the other periods.

Several words have relatively high absolute contributions. For the left part of the plot, these terms are: 'inkomensafhankelijk', 'zorgpremie', 'deelakkoord' and 'nivellering'. In the upper right corner, representing the first and second quarters of 2012, the following terms are the important ones: 'zorg', 'woonakkoord', 'strafbaarstelling', 'toeslagenfraude', 'fyra', 'troonswisseling' and 'inhuldiging'. The lower part of the plot has mostly terms associated with the 2012 campaign, which are: 'premiersdebat', 'langstudeerboete', 'dead' and 'verkenner'. Here, the term 'dead'

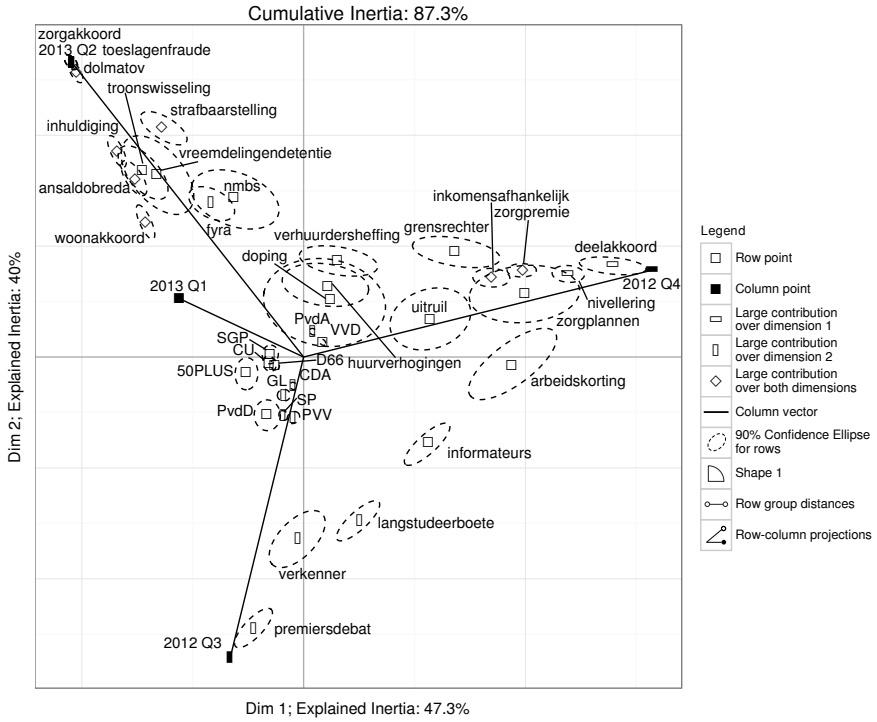

Figure 4: Correspondence analysis solution with row principal normalisation. The solution is calculated from the frequency table of the occurrences of terms and political parties (rows) within the four quarters over the period July 2012 until June 2013 (columns). For this solution, the RMSE is 0.9774.

comes from the phrase 'over my dead body', said by the leader of the SP during the premier's debate on television.

The confidence ellipses around the terms sometimes overlap each other. This indicates similarity in the periods that these terms are mentioned. These could be related terms, but sometimes also unrelated terms. For example, the terms 'doping' and 'huurverhogingen' are both unrelated terms over the period from the end of 2012 until into 2013. However, both these terms were used in the news in the beginning of March 2013. This explains their similarity.

For terms that are related to events with overlap between seasons, the CA plot will show them in between the seasons. Closer to one season means that it is mentioned relatively more often in that season. This leads to a path from each season to the other and can be interpreted as an approximation of the time of the event. It provides a clear overview of when each term was significant. In addition, when it also gravitates to another period, the relevancy towards that period is shown.

4.3.3 Terms by parties over the whole period

In Figure 5 the CA of the co-occurrence of terms (rows) with the political parties (columns) in the period July 2012 until June 2013 is shown. In this case $\alpha = \frac{1}{2}$ in Equation 18 so that the terms can be related to the political parties, and vice versa, using the biplot.

The first dimension on the horizontal axis mostly distinguishes between cooperating parties on the left and coalition or non-cooperating parties on the right. In the right half there is a clear distinction between the winners and losers of the premier's debate during the election period. The SP and PVV were the ones that lost votes, where the VVD and PvdA gained them and thereby formed a ruling coalition.

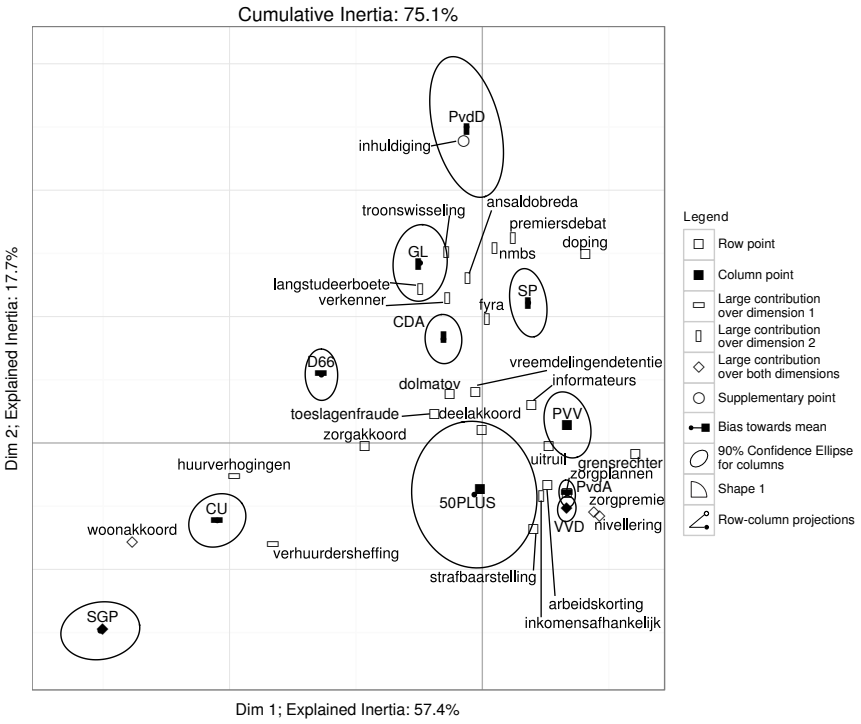

Figure 5: Correspondence analysis solution with symmetric normalisation. The solution is calculated from the frequency table of the co-occurrences of terms (rows) with political parties (columns) in the Dutch newspapers. For this solution, the RMSE is 0.4812.

The cooperating parties are located mostly on the right. This is especially about the 'woonakkoord', which is an agreement made by the coalition with SGP, ChristenUnie and D66. Of these three parties, D66 is located further away from the other two Christian parties.

The D66 tends to also concern itself with other topics, which are located in the upper part of the plot. Here are terms mentioned throughout the year which have been brought to the attention by several other parties in the opposition ('langstudeerboete', 'troonswisseling', 'verkenner').

This is where mostly left-wing parties like PvdD, GroenLinks and SP are located. Also the CDA has interest in these topics.

The PVV has its own spot among the two terms 'eurogroep'and 'grensrechter'. This party has among all words in the plot focused mostly on opposing the euro and discussing the beating of the football linesman ('grensrechter'). 50PLUS has a large confidence ellipse, with less proof that its shown position is the correct one.

The PvdA and VVD, having been in the ruling coalition together for most of the period, are close together in this plot. The important words here are 'inkomensafhankelijk', 'zorgpremie' and 'nivellering'. This concerns the discussion about the plans of the cabinet at the beginning of their ruling period. Most of the words in this region are about opinion differences between the two parties.

4.3.4 Terms by parties per quarter

The CA for each quarter gives insight into the movements of parties relative to each other over time. From these results I found that a distinction can be made between larger and smaller parties. Moreover, there is a clear difference between financial and 'green, humanistic' themes.

I calculated the CA solution for each quarter using principal coordinates for both rows and columns. This results in plots with principal normalisation, making them more readable, but also allowing the distances between the points to be interpretable. This is important, because I examine the movements of the column points (parties), and the differences between them, over time.

Using Procrustes rotation, I rotated each CA solution so that the column points (parties) were closest to the column points of the previous quarter. Since the third quarter of 2012 is the first period here, I rotated this CA plot with the CA solution of the full year (Figure 5) as the basis.

To find out if there is a difference between each period, I calculated the congruence coefficient of the CA coordinates of each quarter with its previous quarter. See Table 4 for these results. Each of the periods has some similarity to the previous period, but not very high. This means that after the rotations, it is possible to compare the CA solutions of each quarter with each other. In addition, there is also some movement of the parties' positions in the CA plot.

Period	Congruence Coefficient
2012 Q3	0.812
2012 Q4	0.874
2013 Q1	0.820
2013 Q2	0.845

Table 4: The congruence coefficients comparing the CA coordinates of each quarter with its previous quarter. The CA coordinates of the 3rd quarter of 2012 is compared with the CA coordinates of the full year from Figure 5.

Within the resulting separate plots (see appendix D) sometimes an outlier was found as one with high absolute contribution, but having few observations. These outliers were taken as supplementary points.

Each of the plots are difficult to interpret together. Therefore, in Figure 6 I have plotted the chronological path of the parties with the party names at the endpoints. From this plot one can derive that within this time period some parties stay within one region, while others move about more. When taking into account the terms positioned in the plots and the position of the parties, an interpretation can be made. The right part contains larger parties with influence, while the left part are mostly parties from the opposition. Also examining the terms in the original plots in appendix D, the upper quadrant contains green and humanistic themes (e.g. 'gaswinning', 'strafbaarstelling', 'vreemdelingendetentie'). The lower quadrant is more focused on financial themes (e.g. 'toeslagenfraude', 'verhuurdersheffing', 'middeninkomens').

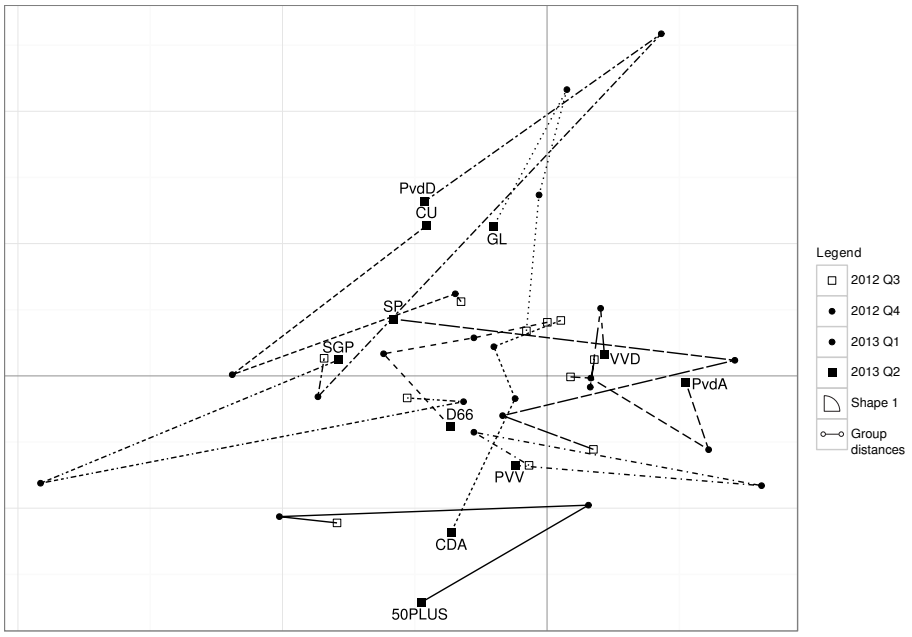

Figure 6: Path of the CA column points (political parties) over the quarters in the period July 2012 until June 2013 (see Appendix D for each single plot). Each quarter's solution is calculated from the frequency table of the co-occurrences of terms (rows) with political parties (columns) in the Dutch newspapers within that quarter. Orthogonal Procrustes rotation is applied on each quarter with reference to its previous quarter.

The MSE for each row point and the row, column and total RMSE are shown in Table 5. This table shows how stable each of the row points are and gives extra information on the uncertainty

of the points in plot of Figure 6. The total RMSE shows that the the third quarter of 2012 is overall the least stable, while the first quarter of 2013 is the most stable. Looking at the MSEs, especially the row point 50PLUS has the largest instability measure for most periods, except 2013 Q1. Overall, the most stable points are the PvdA and VVD.

MSE of row points	2012 Q3	2012 Q4	2013 Q1	2013 Q2
PvdA	0.06	0.23	0.32	0.48
VVD	0.32	0.25	0.19	0.10
CDA	0.49	0.56	0.03	1.46
PVV	1.24	1.66	0.72	0.44
D66	0.42	1.00	0.13	0.36
GL	0.33	8.96	3.04	1.33
SP	1.02	0.70	0.18	0.81
CU	1.67	2.74	0.42	1.69
50PLUS	8.44	14.17	0.60	2.54
PvdD	6.01	6.34	4.26	2.01
SGP	2.48	1.12	1.46	1.12
RMSE of rows	0.4791	0.3292	0.2499	0.2998
RMSE of columns	0.4341	0.4651	0.2771	0.3318
Overall RMSE	0.4566	0.3972	0.2635	0.3158

Table 5: Stability measures for each quarter in the period July 2012 until June 2013. The mean squared error (MSE) of the rows, the relative mean squared error (RMSE) for both rows and columns, and the overall RMSE are shown.

The right hand side of the plot is dominated by the VVD and PvdA. In addition, it is visited by PVV and SP in the first quarter, and together with 50PLUS also in the third quarter. PVV and SP are in the first quarter indeed large parties in the polls, together with VVD and PvdA. Also, 50PLUS is peaking in the third quarter of this period (source: Peil.nl).

Groenlinks is only positioned in the top-left quadrant of the plot. ChristenUnie and PvdD are the two parties that are moving through here as well in addition to moving to the lower part. These are indeed parties with green and humanistic themes.

Where other parties mostly stay around in their place in the plot, the SP moves around the plot a lot towards different places. Also, CDA and D66 don't have their own position throughout the period, moving from the upper part to the bottom of the graph. This could show that over time these parties change their stance towards different themes.

The SGP, 50PLUS and PVV each have their own position which they revolve around. These are parties mostly concerning only certain subjects in the news. For example, 50PLUS is concerned more with the pensions and care of older people, and the PVV is concerned with immigration and foreigners. This is reflected in the plots for each quarter, for these plots see appendix D.

What is interesting, is that the CDA has in the last quarter moved towards 50PLUS and away from other parties. This can be explained by their changing stance towards aiding people of age

50 and higher (source: Aangenomen resoluties CDA Partijcongres van 1 juni 2013). Moreover, the changing position of SP is interesting to note as it shows a changing portrayal of this party by the newspapers.

5 Conclusion

The research question of this thesis was:

How do Dutch national newspapers portray the different Dutch political parties and their relationships?

In this section, I will answer this question with my main findings, showing the relationships and assocations between terms used by newspapers and the political parties. Also, I describe the important terms and themes used for the parties and the analysis over time. Finally, I explain how CA has helped me to arrive at these findings, the limitations that I faced and possible future research on this topic.

I have found that throughout the newspapers, divisions exist between the mentioned terms and parties. Mainly between coalition and opposition parties, but also between green, humanistic and financial themes. The different newspapers each write more on some parties than others. Like the Christian newspapers write more on Christian parties as opposed to other newspapers. In addition, 'de Telegraaf', 'NRC Handelsblad', 'NRC.NEXT' and 'Financieel Dagblad' write more about the coalition parties than others.

Within the terms that I found to be relevant for each period, several terms were identified as being important for different political parties. The three parties D66, ChristenUnie and SGP are mentioned in newspapers in conjunction with the living agreement ('woonakkoord') and raise of housing rent ('huurverhogingen'). The VVD and PvdA were mentioned most with the terms care premium ('zorgpremie') and nivellation ('nivellering').

In the long term, I found the green, humanistic and financial themes to be important for the portrayal of parties. GroenLinks, PvdD and ChristenUnie all tend to be mentioned with terms on green and humanistic themes. Whereas 50PLUS, PVV and SGP are mostly associated with financial terms. This differentiation differs per quarter, D66 and CDA both move towards financial themes as time progressed. Through the terms used, I also found relationships between parties. VVD and PvdA are most of the time mentioned together. As time progresses, the CDA has moved their focus into the direction of topics of 50PLUS.

My findings are based on the use of Correspondence Analysis. This method has helped me to analyze the terms used in newspapers and to find the relationships between parties and terms. In addition to this, the Procrustes Rotation method has allowed me to interpret the CA plots over time. Moreover, the method for finding terms associated with events in certain periods provided useful terms for this analysis.

CA is a powerful knowledge discovery tool for text mining. It finds assocations between terms and also aids in mapping a timeline of the events depicted as terms. Using this, it is easy to relate these events to other terms, in this case the political parties. So that also the importance of each of these parties can be examined over time.

Doing this research I faced some limitations to the used methods. When finding the relevant terms, it proved difficult to find a good method. The technique I used has been very useful, as

it was simple and effective. However, it still required tweaking by removing certain terms.

As CA is an explorative technique, this research could lead to a more in-depth analysis of the terms used by newspapers. For future research one can for example look into the sentiment associated with each term. Also, one can have a closer look on the relationship between the associated terms and the political polls.

As I have shown in this thesis, it is indeed possible to apply Correspondence Analysis on text data to show relationships and assocations between terms. It is easy to think how this can be incorporated with other text sources, topics, newspapers or languages.

References

Aangenomen resoluties CDA Partijcongres van 1 juni 2013. URL `http://www.cda.nl/partijcongres/`.

Hervé Abdi. Rv coefficient and congruence coefficient. *Encyclopedia of measurement and statistics*, pages 849–853, 2007.

Jean-Paul Benzécri. *L'analyse des données: l'analyse des correspondances*, volume 2. Dunod, 1973.

Adam Bermingham and A.F. Smeaton. On using Twitter to monitor political sentiment and predict election results. 2011.

Ingwer Borg and Patrick J.F. Groenen. *Modern Multidimensional Scaling*. Springer, second edition, 2005.

Neal Caren. Cleaning up LexisNexis Files, May 15 2012. URL `http://nealcaren.web.unc.edu/cleaning-up-lexisnexis-files/`.

M.D. Conover and B. Gonçalves. Predicting the political alignment of twitter users. *Privacy, security, risk and trust*, 2011.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Miles Efron. Using cocitation information to estimate political orientation in web documents. *Knowledge and Information Systems*, 9(4):492–511, September 2005.

John C. Gower, Patrick Groenen, Michel van de Velden, and Karen Vines. Better perceptual maps: Introducing explanatory icons to facilitate interpretation. *Food Quality and Preference*, 2014 (in press).

Michael J. Greenacre. *Theory and applications of correspondence analysis*. 1984.

Michael J. Greenacre. *Correspondence Analysis in Practice*. 1993.

Michael J. Greenacre. *Biplots in Practice*. 2010.

Steven Grijzenhout, Valentin Jijkoun, and Maarten Marx. Opinion mining in Dutch Hansards. *Proceedings Workshop From Text to Political Positions*, pages 1–15, 2010.

HOI, Instituut voor Media Auditing. URL `http://hoi-offline.staging.modernmedia.nl`.

John R. Hurley and Raymond B. Cattell. The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7(2):258–262, 1962.

Jan Kleinnijenhuis, Dirk Oegema, Jan de Ridder, Anita van Hoof, and Rens Vliegenthart. *De puinhopen in het nieuws*. 2003.

Walter Kristof. A theorem on the trace of certain matrix products and some applications. *Journal of Mathematical Psychology*, 7(3):515–530, 1970.

LexisNexis Academic. URL `http://www.lexis-nexis.com`.

Urbano Lorenzo-Seva and Jos M.F. Ten Berge. Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(2):57–64, 2006.

NOS. Wilders: glorieuze dag voor Nederland, June 10 2010.

Peil.nl. URL `http://panel.noties.nl/peil.nl/`.

Trevor J. Ringrose. Alternative confidence regions for canonical variate analysis. *Biometrika*, 83 (3):575–587, 1996.

Arno Scharl and Christian Bauer. Mining large samples of web-based corpora. *Knowledge-Based Systems*, 17(5-6):229–233, August 2004.

Ryuma Shineha, Aiko Hibino, and Kazuto Kato. Analysis of Japanese newspaper articles on genetic modification. *Journal of Science Communication*, 7(2):1–9, 2008.

Tweede Kamer der Staten-Generaal. URL `http://www.tweedekamer.nl`.

Michel van de Velden and Henk A.L. Kiers. Rotation in correspondence analysis. *Journal of classification*, 22(2):251–271, 2005.

Michel van de Velden, Alain de Beuckelaer, Patrick J.F. Groenen, and Frank M.T.A. Busing. Solving degeneracy and stability in nonmetric unfolding. *Food Quality and Preference*, 2012.

Volkskrant. Peiling: SP groeit naar 37 zetels, August 12 2012.

Artur Šilić, Annie Morin, Jean-Hugues Chauchat, and Bojana Dalbelo Bašić. Visualization of temporal text collections based on Correspondence Analysis. *Expert Systems with Applications*, 39(15):12143–12157, November 2012.

Cornelia Zuell. Using computer-assisted text analysis to identify media reported events. *Social Science Computer Review*, 26(4):483–497, 2008.

A Dutch stopwords

aan	dus	iets	niets	van	zich
af	een	ik	nog	veel	zij
al	eens	in	nu	voor	zijn
alles	en	is	of	want	zo
als	er	ja	om	waren	zonder
altijd	ge	je	omdat	was	zou
andere	geen	kan	onder	wat	
ben	geweest	kon	ons	we	
bij	haar	kunnen	ook	wel	
daar	had	maar	op	werd	
dan	heb	me	over	wezen	
dat	hebben	meer	reeds	wie	
de	heeft	men	te	wij	
der	hem	met	tegen	wil	
deze	het	mij	toch	worden	
die	hier	mijn	toen	wordt	
dit	hij	moet	tot	zal	
doch	hoe	na	u	ze	
doen	hun	naar	uit	zei	
door	iemand	niet	uw	zelf	

B Extracted terms for each period

2012-2013	2012 Q3	2012 Q4	2013 Q1	2013 Q2
ansaldobreda	body	arbeidskorting	aardbeving	afscheidsbrief
arbeidskorting	campagneteam	armstrong	aardolie	akkoorden
deelakkoord	campagnetijd	belastingschijven	achtergestelde	ansaldobreda
dolmatov	dead	belastingtarieven	ansaldobreda	bangladesh
doping	doorrekeningen	belastingverlaging	bankverzekeraar	begrotingsnorm
fyra	flanken	beëdiging	bouwend	bendes
grensrechter	ipsos	bordes	cameron	bulgaren
huurverhogingen	kandidaatkamerleden	deelakkoord	cito	detentie
informateurs	kieskompas	formatieonderhandelingen	cyprus	detentiecentrum
inhuldiging	langstudeerboete	grensrechter	eurogroep	dolmatov
inkomensafhankelijk	lijsttrekkersdebat	heffingskorting	eurogroepvoorzitter	enkelband
langstudeerboete	middenkabinet	informateurs	fyra	examens
nivellering	nekaannekrace	inkomensafhankelijk	gaswinning	feitenrelaas
nmbs	opiniepeilers	inkomensverschillen	groningers	fiod
premiersdebat	premiersdebat	kiezersbedrog	huurverhogingen	fyra
strafbaarstelling	rtl4	koopkrachtdaling	hypotheekregels	gevangeniswezen
toeslagenfraude	stemhokje	koopkrachtplaatjes	inhuldiging	humaner
troonswisseling	stemwijzer	middeninkomens	keulen	inhuldiging
uitruil	tns	ministersposten	mijnen	nmbs
verhuurdersheffing	tvdebat	modaal	nationalisatie	opsluiting
verkenner	tvdebatten	nibud	neuroloog	pensioenopbouw
vreemdelingendetentie	tweestrijd	nivellering	polderoverleg	schaliegas
woonakkoord	verkenner	oproer	reaal	strafbaarstelling
zorgakkoord	verkiezingsavond	radiostilte	richter	toeslagen
zorgplannen	verkiezingsdag	regeringsverklaring	sns	toeslagenfraude
zorgpremie	verkiezingsdebatten	tentenkamp	spaarders	troonswisseling
	verkiezingsretoriek	uitruil	troonswisseling	vreemdelingenbeleid
	verkiezingsstrijd	uitschieters	verheffing	vreemdelingendetentie
	verkiezingstijd	zorgplannen	verhuurdersheffing	winning
	zwevende	zorgpremie	woonakkoord	zorgakkoord

C Python script for extracting articles from LexisNexis article list

```
#!/usr/bin/env python
# encoding: utf-8
"""
split_ln.py

Created by Neal Caren on 2012-05-14.
neal.caren@unc.edu

Takes a downloaded plain text LexisNexis file and
converts it into a CSV file.


sample usage:
£ python split_ln.py T*.txt
Processing The_New_York_Times_TP_2012_1.txt
Processing The_New_York_Times_TP_2012_2.txt
Done

£ python split_ln.py ap_tp_201201.txt
Processing ap_tp_201201.txt
Done

"""



def split_ln(fname):
    print 'Processing\t',fname
    #Imort the two required modules
    import re
    import csv
    outname=fname.replace(fname.split('.')[-1],'csv') #replace the
                                        #extension with "csv"
    #setup the output file. Maybe give the option for seperate
    #text files, if desired.
    outfile=open(outname,'wb')
```

```
writer = csv.writer(outfile)

lnraw=open(fname).read() #read the file

workfile=re.sub('      Copyright .*\\n','ENDOFILE',lnraw)
#silly hack to find the end of the documents
workfile=workfile.replace('\xef\xbb\xbf\n','') #clean up crud
                          #at the beginning of the file
workfile=workfile.split('ENDOFILE') #split the file into a list
                                    #of documents.
workfile=[f for f in workfile if len(f.split('\n\n'))>2]
#remove an blank rows

#Figure out what special meta data is being reported
meta_list=list(set(re.findall(
'\\n([A-Z][A-Z][A-Z][A-Z][A-Z-]*?):',lnraw))) #Find them all
exceptionList=('CU','NS','CDA','SGP','PVV','EU','SP','VVD',
    'GL','KNMI','NAVO','ANWB')
meta_list=[m for m in meta_list if float(
    lnraw.count(m))/len(workfile)>0.2 and
    m not in exceptionList] #Keep only the commonly occuring
                            #ones
meta_tuple=('SEARCH_ROW','PUBLICATION','DATE','TITLE','EDITION')
for item in meta_list:
    meta_tuple=meta_tuple+(item,)
writer.writerow(meta_tuple+('TEXT',))

#Begin loop over each file
for f in workfile:

    #Split into lines, and clean up the hard returns at the end
    #of each line. Also removes blank lines that the occasional
    #copyright lines
    filessplit=[row.replace('\n',' ') for row in f.split(
        '\n\n') if len(row)>0 and 'All Rights Reserved' not
        in row]
    #print filessplit[0:10]
    #The id number (from that search) is the first text in the
    #first item of the list
```

```
            docid=filessplit[0].lstrip().split(' ')[0]
            dateedition=filessplit[2].lstrip()
            date=dateedition.split(' ')[0]+' '+dateedition.split(
                ' ')[1]+' '+dateedition.split(' ')[2].replace(',','')
            edition= dateedition.replace(date,'').split(
                '                          ')[-1].lstrip()
            if 'GMT' in edition:
                edition=''
            title=filessplit[3]
            publication=filessplit[1].lstrip()
            #Extra the text and other information
            text=''
            meta_dict={k : '' for k in meta_list}
            for line in filessplit:
                if len(line)>0 and line[:2]!='  ' and line!=
                        line.upper() and len(re.findall('^[A-Z]
                        [A-Z-]*?:',line))==0 and title not in line:
                    text=text.lstrip()+'\n'+line.replace('","',' , "')
                else:
                    metacheck=re.findall('^([A-Z][A-Z-]*?):',line)
                    if len(metacheck)>0:
                        if metacheck[0] in meta_list:
                            meta_dict[metacheck[0]]=
                                line.replace(metacheck[0]+': ','')
            if len(re.findall(
                    '^[A-Z][A-Z][A-Z][A-Z][A-Z][A-Z-]*?:',title))>0:
                title=''
            #Output the results to a csv file
            meta_tuple=(docid,publication,date,title,edition)
            for item in meta_list:
                meta_tuple=meta_tuple+(meta_dict[item],)
            writer.writerow(meta_tuple+(text,))
            #output.write(docid+'\t'+title+'\t'+text+'\n')
        print 'Wrote\t\t',outname


if __name__ == "__main__":
    import sys
    try:
```

```
    flist=sys.argv[1:]
except:
    print 'Only one argument please. But you can use
            things like *.txt'
else:
    for fname in flist:
        split_ln(fname)
    print 'Done'
```

D Co-occurences per quarter

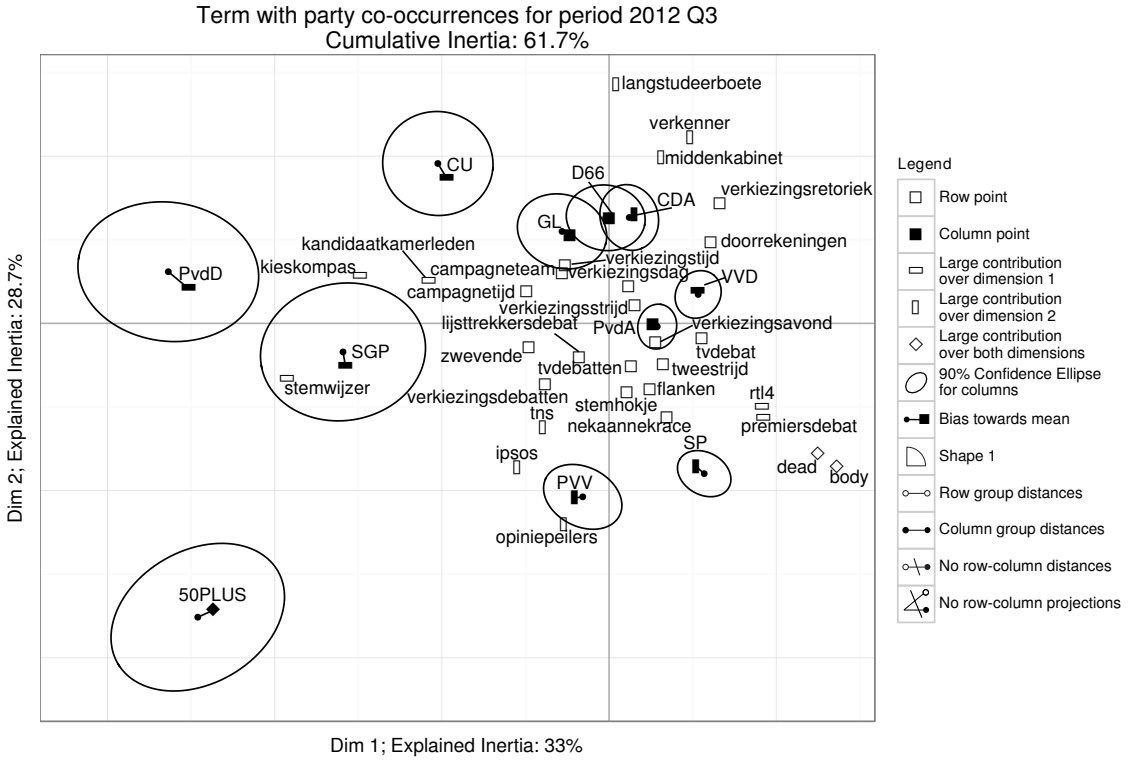

Figure 7: Correspondence analysis solution with principal normalisation. The solution is calculated from the frequency table of the co-occurrences of terms (rows) with political parties (columns) in the Dutch newspapers.

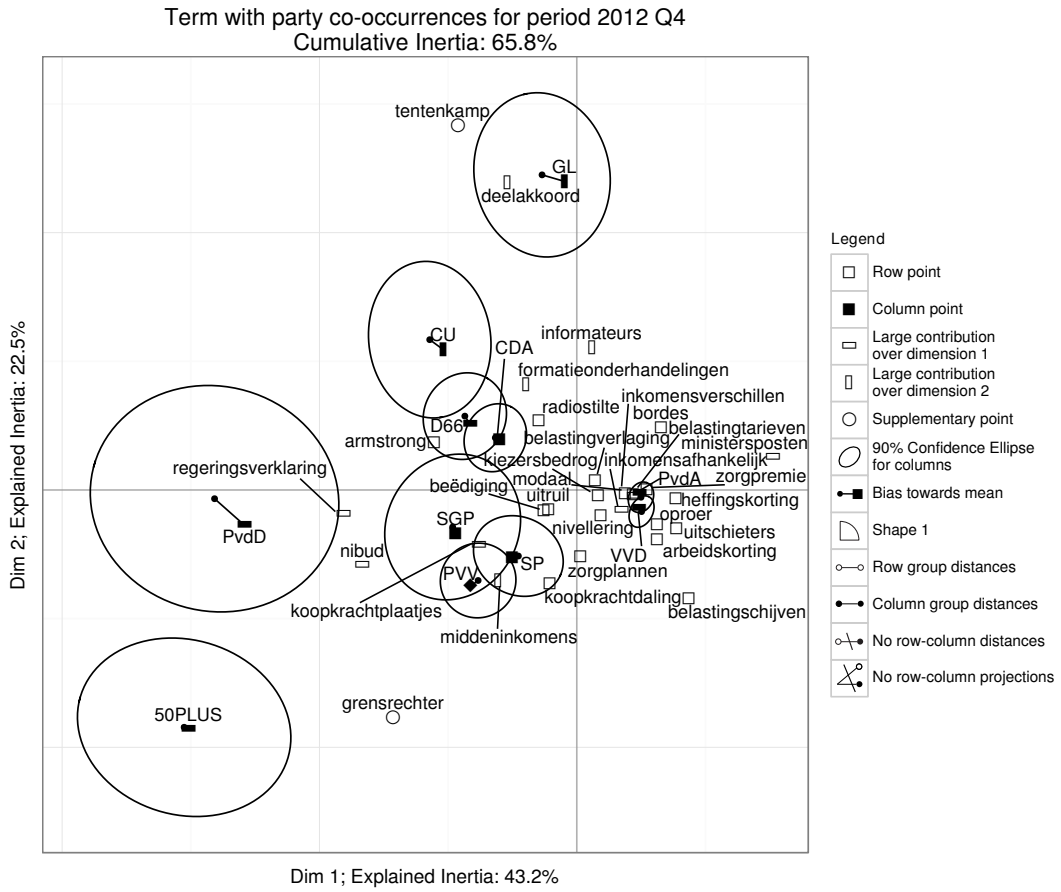

Figure 8: Correspondence analysis solution with principal normalisation. The solution is calculated from the frequency table of the co-occurrences of terms (rows) with political parties (columns) in the Dutch newspapers.

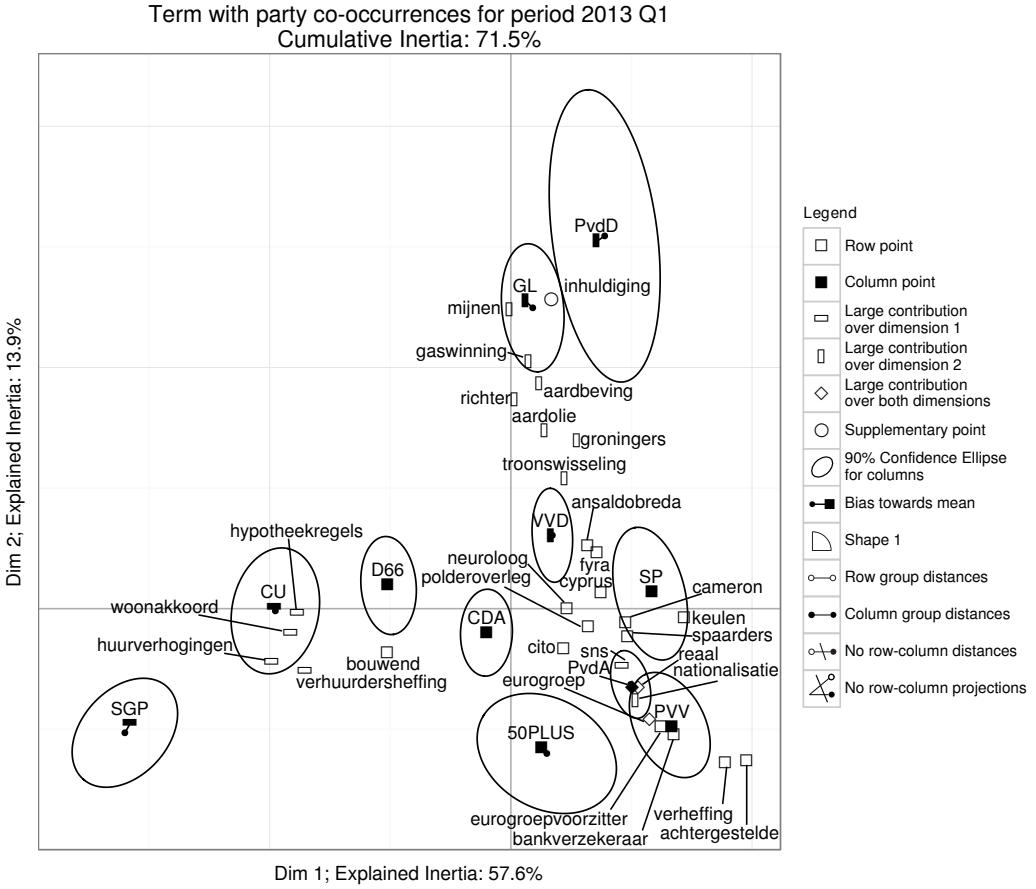

Figure 9: Correspondence analysis solution with principal normalisation. The solution is calculated from the frequency table of the co-occurrences of terms (rows) with political parties (columns) in the Dutch newspapers.

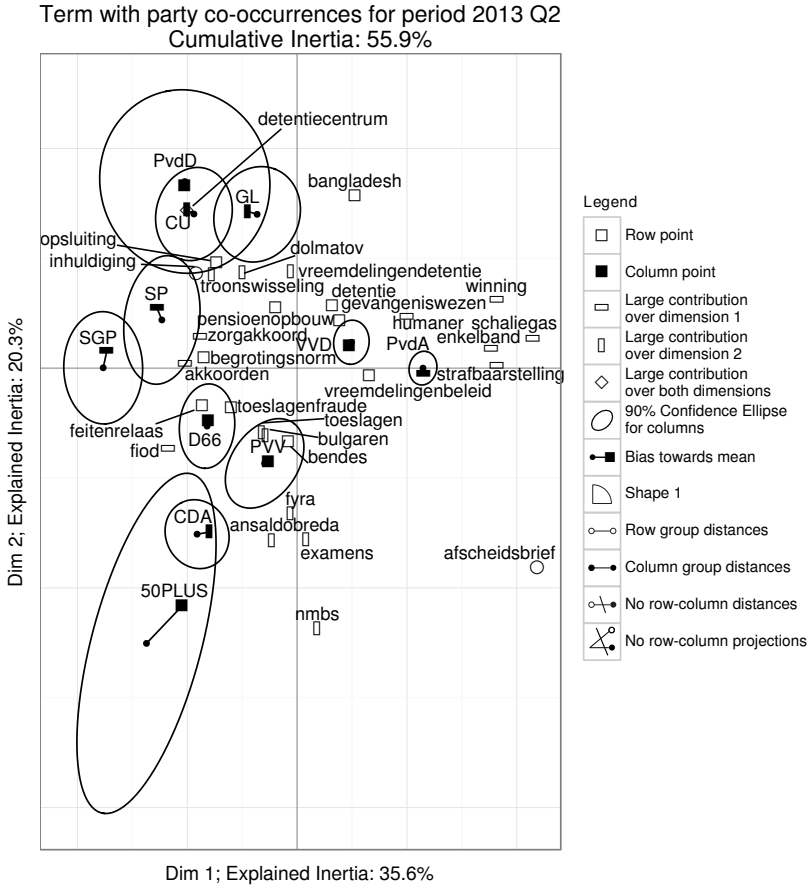

Figure 10: Correspondence analysis solution with principal normalisation. The solution is calculated from the frequency table of the co-occurrences of terms (rows) with political parties (columns) in the Dutch newspapers.