Text Mining

Post-electoral survey in Belgium in 2009

A. El Khaloui - M. Le Tertre - V. Moulinier

Contents

List of Figures

1 Data presentation and main issue of our study

1.1 PIOP data base

The data we used for our study come from a post-electoral study led in Belgium by PIOP. This research aims to know and to understand better electoral choices that have been made. Thus, the participants answered several questions:

- Closed questions:

  Votes at the House of Representatives between 1987 and 1991.

  Self-positioning on a « Left-Right » scale (as it would seem, often misunderstood)

  Social and demographic information

- Open questions:

  What is your opinion on the « Left » party?

  What is your opinion on the « right » party?

1.2 Purpose of the PIOP study

For our study, we decided to focus on the raw words on not on the lema. Thus, our objectives for the study are:

- Determine the different quoted themes and their relative significance for each of the two open questions. What are the main conclusions we can draw from the words and segments lists?
- Compare the lists of words that have been used to characterize, respectively, the « Left » and the « Right » parties.
- Study, with the help of a discrete analysis, the answers. Try to determine the closed questions linked to the open answers.
- Study simultaneously answers to both questions using a MFA on contingency tables.
- Classify the participants regarding their factorial coordinates returned by the MFA. Characterize the clusters and give meaning to the results.
- Synthesis and conclusion

1.3 Quick overview of the data

This study aims to deal with a political matter, since then we want to know what our population is « made of ». Giving a quick loog at the summary of our data in R, we can see several things:

- The gender is quite balanced (680 women and 743 men)
- We have a balanced distribution of our participants into the three main social and professional categories.
- The age is also well distributed among our population
- Concerning the education level is not homogeneous since the majority of people followed ...

1.4 Corpus information

Our corpus provides 21760 occurences on 2129 different words provided by 1423 participants.

2 Main themes for each party

2.1 The « Left »

The « Left » is described with 11614 occurences on 1481 words. Using the TextoMineR package we manage to get the contingency table providing a table crossing our participants with the words they used. We take the *stopwords* out this table (1402 words remaining). We decide to consider a threshold of 10 for the minimal number of documents and word frequencies (96 words remaining).Also, we take out of our study the « I don?t know » kind of words (90 words remaining).

2.1.1 CA analysis and interpretation for the « Left »

As we can notice on the figure 1, the eigenvalues for the first 10 dimensions are higher than 0.5 meaning we are dealing with a « dichotomy », which can be observed on the figures 2 and 3, i.e. we can see three definite groups.

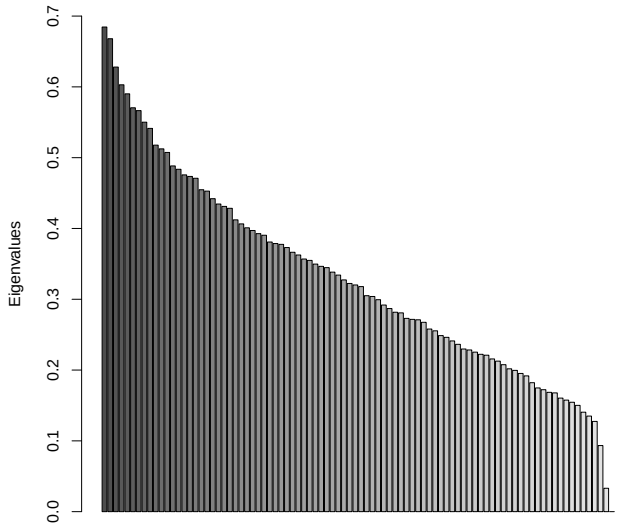

Figure 1: Eigenvalues of the CA performed on the « Left » question

Those three cores are oppositions in term of description concerning the « Left » party. After performing the Correspondence Analysis, we proceed to observing the factor and individuals maps:

- The individuals map shows three distinct groups of respondents, drawing a triangular shape. We can hence assume that there is three types of perception concerning the left parties.

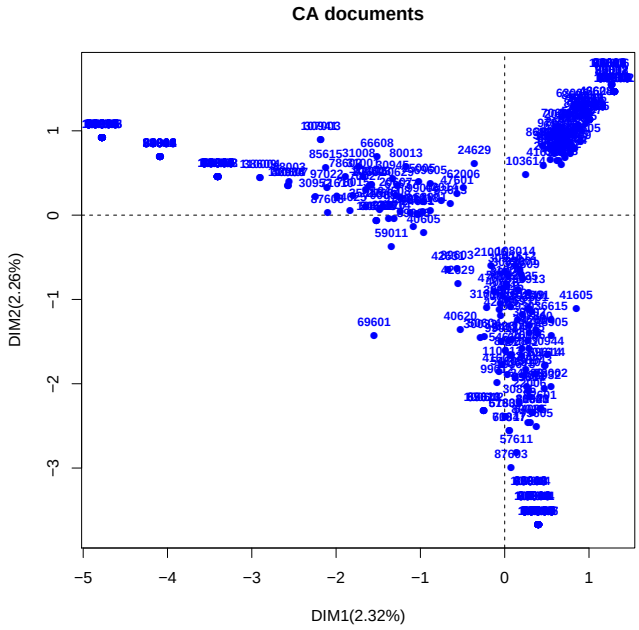

Figure 2: CA factor map of individuals on the « Left » question

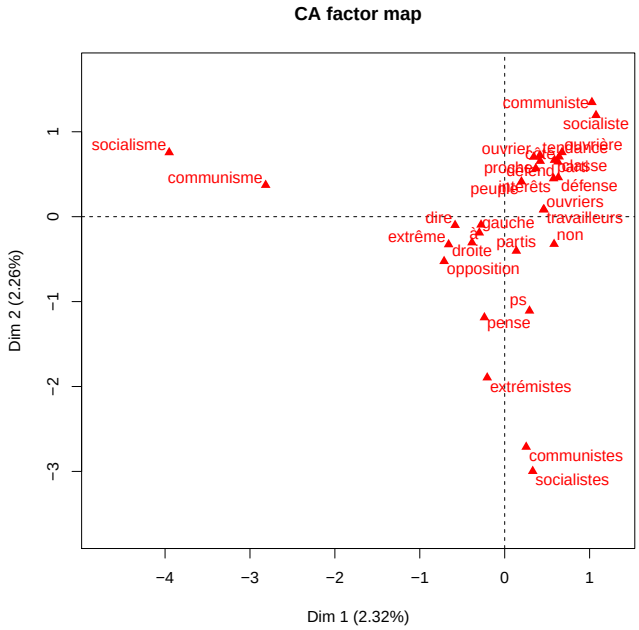

Figure 3: CA factor map of words on the « Left » question

- Examining the factor map gives us the elements to explain how the three groups are constituted.

  The first group (top right) contains the people who see the left as a very political concept, they use politics-related terms like "peuple", "travailleurs", "opposition", "droite", "défense", "partis", "gauche". These people are probably politically engaged and have a certain knowledge of politics. The majority of the respondents belongs to this group, thus we can assume that a lot of people here have a very mainstream political culture, shaped mainly by mass media and common talks.

  The second group (top left) is made out of the people who see the left as an ideology, a doctrine and probably a historical reference. They use the words "communisme" and "socialisme"

  The third group seems (bottom right) to have a more personal vision of the left. They describe it as a group of people whom they call "extrémistes", "communistes" and "socialistes". The presence of the word "pense" makes it more obvious that these people consider politics as a very personal concept, linked to individual choices.

- The previous information can be found in the metakeys output, concerning the two first principal components. We can also find more in the description of superior components. For instance:

  The positive side of Dim3 contains people who think that the left/right opposition is an obsolete idea that should no longer be considered for it is a nonsense. They use words like "mauvais", "opposition", "plus", "moins", "sens".

  The positive side of Dim4 has a very sociological vision of the Left, that seems to be related to "Class Conflict" and Marx theories about industrialized societies. They use words like "mauvais", "ouvrière", "classe", "côté", "défense".

2.1.2 Deeper analysis for the « Left »

- We run the textual analysis again about the « Left », this time with aggregating age and education within a single categorical variable. The main goal here is to see wether or not this aggregation creates a partitionning in terms of opinion, assessed via the words used as answers to the open questions.

  If we observe the factor map concerning the `age_education` variable, we observe an age gradient going from the top of Dim2 to its bottom.

  There is also a slight trend in terms of education, going from the left to the right of Dim1 (cf.figure 4). Note that the "De 18 a 24 anys,sans études" subsample is isolated on the factor map with few words around it, on account of its size. Indeed, it is quite hard to find completely uneducated young people in 2009's Belgium and it is undoubtedly a very specific population that consists of people whom dropped out school for peculiar social causes.

- The same results are obtained when the analysis is runned on the question about the Right.

- We now in light of the foregoing observe the words map, made with the fourty terms that most contribute to constituting it. The first thing we notice is that the question we ask is perceived as difficult, what explains the fact that a lot of people just can't give their opinion.

  The "under educated" people use words like "je", 'ne", "sais", "aucune", "pas", "rien". They don't have a precise idea about what the Left is and are totally OK with saying so.

The more educated people on the other hand all have an opinion about the question. This can be linked to education but also to some forme of pride. They use political lexical field: "groupe", "sens", "société", "sociales", "intérêts".

Concerning age categories, there is no particular patterns.

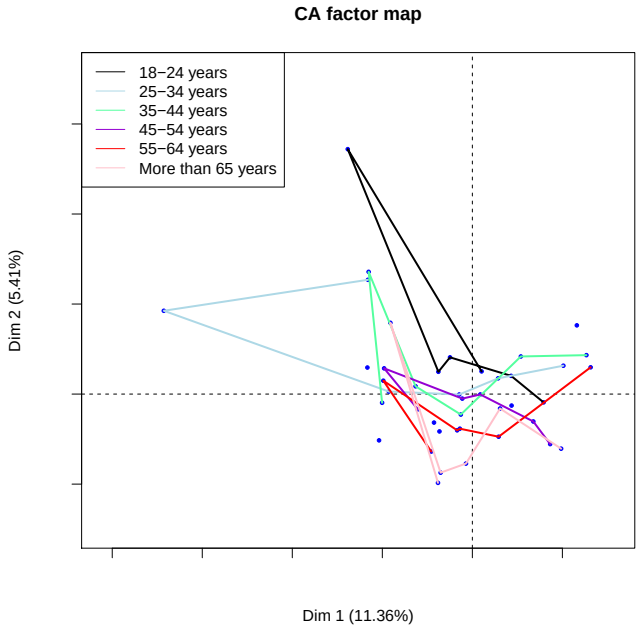

Figure 4: CA factor map of individuals aggregated on the « Left » question

2.2 The « Right »

The « Right » is described with 10146 occurences on 1445 words. Using the TextoMineR package we manage to get the contingency table providing a table crossing our participants with the words they used. We take the *stopwords* out this table (1371 words remaining). We decide to consider a threshold of 10 for the minimal number of documents and word frequencies (107 words remaining).Also, we take out of our study the « I don?t know » kind of words (101 words remaining).

2.2.1 CA analysis and interpretation for the « Right »

As we can notice on the figure 5, the eigenvalues for the first 10 dimensions are higher than 0.5 meaning that, as for the « Left », we are dealing with a « dichotomy », which can be observed on the figures 6 and 7, i.e. we can see three definite groups.

The individuals map shows three distinct groups of respondents, drawing a triangular shape as before. We can hence assume that there is three types of perception concerning the « Right » party.Examining the factor map gives us the elements to explain how the three groups are constituted.

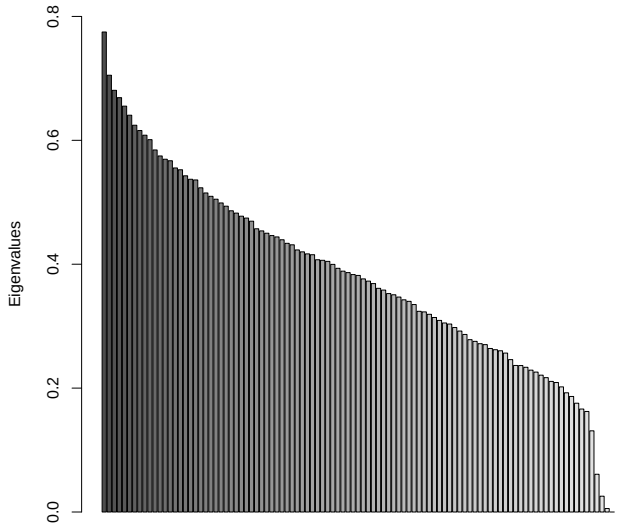

Figure 5: Eigenvalues of the CA performed on the « Right » question

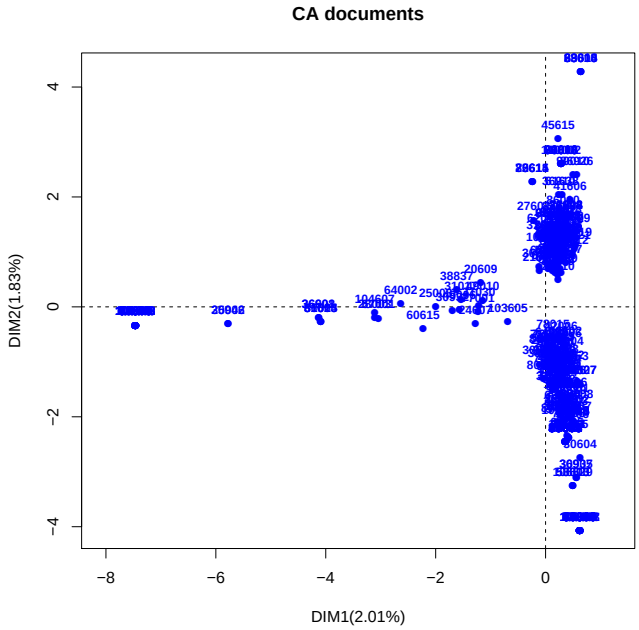

Figure 6: CA factor map of individuals on the « Right » question

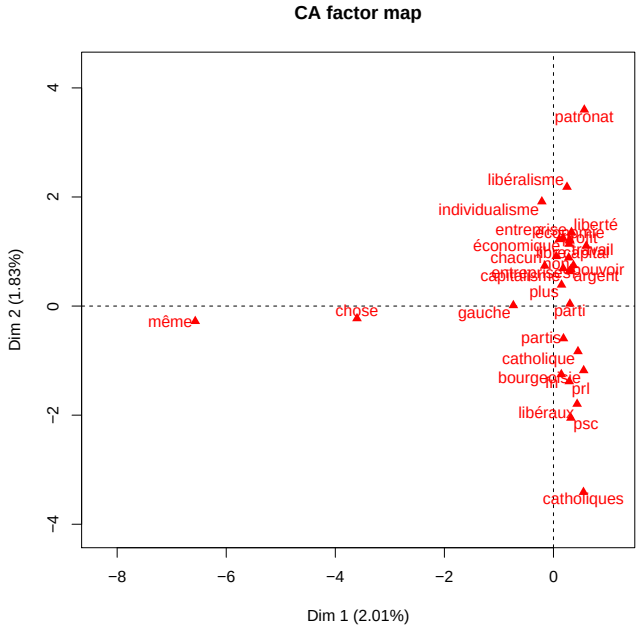

Figure 7: CA factor map of words on the « Right » question

- The first group (top right) gathers the people who see the right as a kind of capitalism in the economic sense of the term. The right corresponds to an american-style of politics, based on the liberalism which defends the free enterprise and the freedom of the market.
- The second group (bottom right) contains the people who see the right as a group of conservatives. They use the words "catholiques", "libéraux", "bourgeoisie". Thus, the second factorial axis split the liberal right of the conservative right.
- The last group (middle left) gathers the people who don't see the difference between the left and the right. They use the words "même", "chose", "gauche".

We can also find more in the description of superior components in the metakeys output. For instance: the negative side of Dim3 and the positive side of Dim4 show the people who also think about the right as extremists

2.2.2 Deeper analysis for the « Right »

As we could see on the figure 7, the first dimension is only constructed with people who claimed that the « Right » is the « same thing as the Left ». In order to get more details about the others, we decide, now, to take out of our table those words (e.g. « même », « chose »). The CA on this table shows a Guttman effect (figure 8). This representation drove us back to the comments we wrote just before. In fact, the Guttman effect shows us an opposition between the participants: opposing extreme people on Dim1 and opposing average to extreme people on Dim2.

- On the left side of the plot we have participants who see the « Right » as people using religious and economic words: the « Right » being the party of people who are « catholiques », « libéraux », « chrétiens», « capitalistes » and those people belonging to the upper class (« bourgeoisie »).
- On the right side of the plot, we have participants who define the « Right » using words of ideology such as «libéralisme », « individualisme », « capitalisme », « liberté », etc. Also those people see the « Right » as a party defending the rights and interests of companies, helping them making benefits.
- At the bottom of the plot we can see people who defined the « Right » as a party characterized with adjectives as « catholique », « conservateur », « extrême droite » and « associale ». Also, those people associated other names of parties such as « PRL », « PSC », « FN », etc.

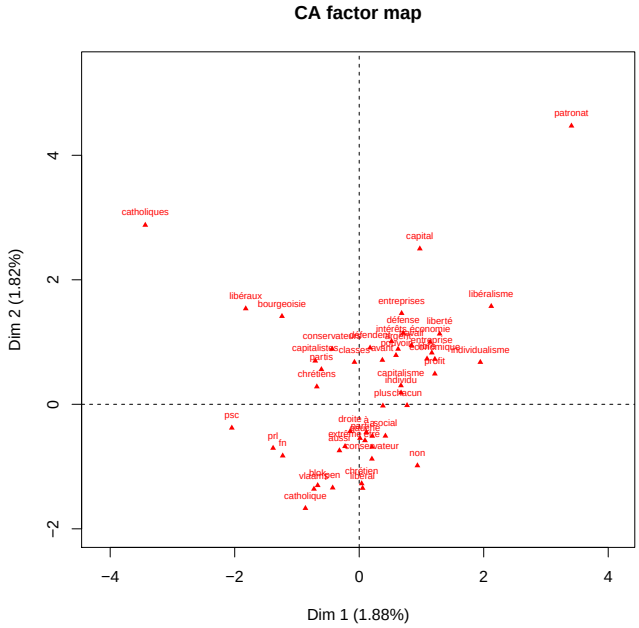

Figure 8: CA factor map of words on the « Right » question - Guttman effect

3 Exploratory analysis of the answers

3.1 Multiple Factor Analysis: exploring simultaneously answers to both questions

We perform an MFA on the contengency table crossing the respondents (as individuals) and the words they used for the two questions about the Right and the Left (as variables). We also use the following variables as supplementary: `age_sexe` , `AgeC` , `études` and `vote`.

The individual factor map shows an ordinary shape of the obtained cloud. The repartition of the individuals on the two first dimensions is equal what let us think that they are both reliable. Moreover, when plotting the

eigenvalues, we see that the first five dimensions are the most meaningful in terms of inertia. This is why we will perform the hierarchical clustering on the five dimensions only.

When observing the Groups representation plot, it appears that they are very close on Dim1 and a bit less on Dim2. Hence, they both induce the same structure on the individuals. In other terms, the answers to the two questions are not that different. Concerning the supplementary variables, they are located at the origin of the map's axis system; which means that they do not influence the opinion of the respondents concerning the two questions.

We also note that the under educated and the elder parts of the sample do not use the same lexical systems for answering the two questions. This is seen thanks to the very distant partial points. On the other hand, the people who attended college are way more consistent in their answers.

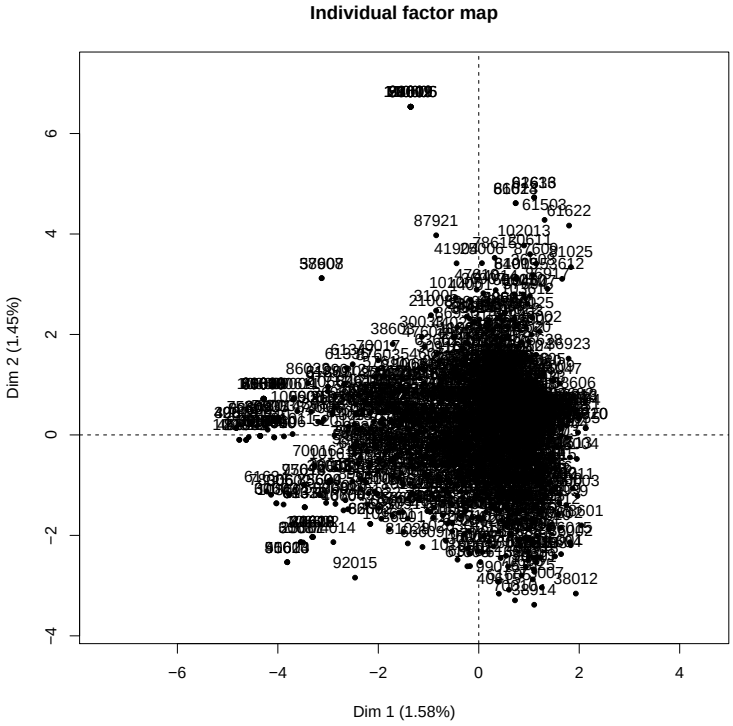

Figure 9: Individual factor map of the MFA

3.2 Clustering analysis on the principal components of the MFA

We perform a HCPC on the five first principal components issued from the MFA with the respondents as individuals in order to classify them. The HCPC function of FactoMineR suggests a partitioning in 5 clusters. Two of them have a small size (clusters 1 and 3).

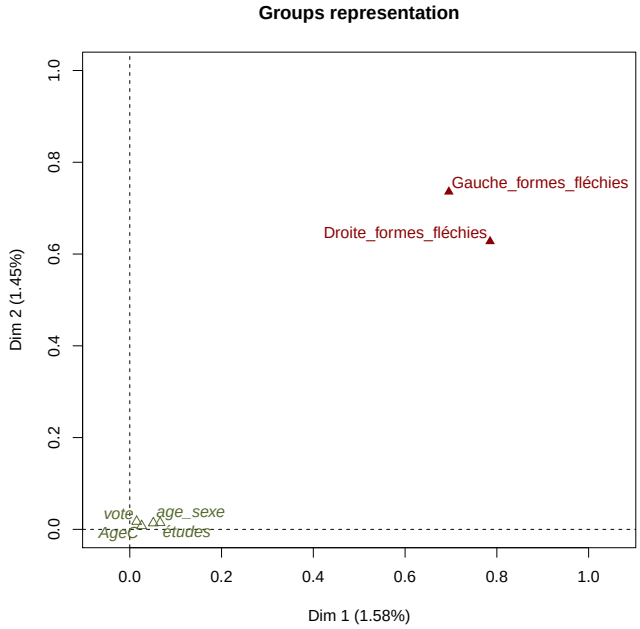

Figure 10: Groups representation of the MFA

Using the `descfreq` function of FactoMineR, we get to describe our clusters with the words they used to give their opinion. The descriptions given for the « Left » and « Right » are consistent with our previous correspondance analysis:

The « Left »

- Cluster 1: They described the « Left » as a party made of « socialistes » and« communistes ».
- Cluster 2: They describe the « Left » with ideology terms such as « socialisme », « communisme » and the party defending the rights of « ouvriers ».
- Cluster 3 cannot be described because they had no opinion.
- Cluster 4: They describe the « Left » as a social situation of the persons voting for it: a party that defends and represents the rights of workers.
- Cluster 5: The « Left » is seen as a political party which is good for the society and defends its interests and has meaningful ideas.

The « Right »

- Cluster 1: They described the « Right » as a party made of people: « catholiques », « chrétiens », « libéraux ». it is associated with real names of parties such as « PSC », « PRL ». Some of the participants think that the « Right » is represented by extremists people.
- Cluster 2: They describe the « Right » with adjectives like: « chrétien », « catholique », « extrême ».

- Cluster 3 cannot be described because they had no opinion.
- Cluster 4: They describe the « Right » as people defending the rights and the interests of companies, helping them to make benefits and be richer. It is seen as party made of people called « capitalistes », « indépendants », « mauvais », « patrons », «riches », etc.
- Cluster 5: The « Right » is seen as a political party which is good for the society and defends its interests and has meaningful ideas: the « Right » is a good party for the economy and it respects the individual rights. Moreover, some of the participants see the people voting for the « Right » as « racistes » and « conservateurs ».

4 Conclusion

To conclude, we will summarize the highlights of this study:

- Depending of their generation, people do not use the same notions to describe the Left and the Right: The young people talk in very personal terms while the old ones have a more global and ideological point of view.
- The asked questions are globally perceived as difficult and a substantial number of respondents are not able to give their opinion. More specifically, the highly educated part of the sample gives an opinion in most of the cases what can also be linked to a form of pride. On the other hand, the least educated say "I don't know" more easily.
- The Right and Left concept are perceived in different ways by people: some see them as political concepts while some other talk only about economical or social notions. It seems that depending on the person's background and interests, they tend to have a different vision.