Starting Points in French Discourse Analysis' Lexicometry to Study Political Tweets

Marge Käsper [0000-0002-0991-4373] and Liina Maurer

University of Tartu, Estonia
marge.kasper@ut.ee
liina.maurer@gmail.com

Abstract. To study a corpus of tweets by the French president Emmanuel Macron (#EmmanuelMacron), the paper presents a series of lexicometry works in French Discourse Analysis that have studied presidential discourse in France. The setup of the analysis of a corpus gathered by the authors is tested and focused by several tools available in Voyant Tools, AntConc and in a special political vocabulary analysis platform. The testing leads to a reasoned presentation of the data organized around one salient French word in this corpus – *devons* ('we have to' in English).

Keywords: Discourse Analysis, Emmanuel Macron, Political Tweets.

1 Introduction

In the Nordic countries, French Language and Culture Studies are probably not the first field one would look at when thinking about digital humanities. However, since in general our study concerns French language context as well, in this paper we propose to take inspiration from French Discourse Analysis, our own analysis aiming to examine a corpus of tweets posted on Twitter by the French president Emmanuel Macron (the account #EmmanuelMacron). Thus, with the background in language and society studies, we will present some starting points in the setup of our analysis. As "at the highest level of generality, one could say that digital humanities designate an interdisciplinary dialogue on the digital dimension of the research in the humanities and social sciences at the level of tools, methods, objects of study and modes of communication"[1], we will discuss these aspects for our analysis. We will explore the tools, various methods and specific objects of study to discuss subsequently, for a socio-discursive perspective, the Twitter as a mode of communication for a president.

The paper is organized as follows. The sections 2 and 3 present the former framework of studies conducted in political lexicometry in France. The sections 4 and 5 present the corpus we have gathered and tested to proceed our research.

2 Discourse Analysis, Political Lexicometry and French Presidents

In Discourse Studies, since the 1960s until today, a part of what is called the French School of Discourse Analysis has been using various machine–based methods to measure the social impact of words in discourse. Since the ideal of an imaginary automatic tool to detect ideology [18] and the first works in political lexicometry at St. Cloud [13], [14], the machine–based methods have thus been discussed, developed and diversified (for these discussions, see [2]), to create various "textometric"[19], "logometric" [15] or "ideometric" [5] analyses. As for the Presidents as object of study, it is interesting to note that already the radio speeches by General de Gaulle from 1958 to 1965 have been studied for their vocabulary statistics although, maybe first of all, the study was important for a "systematic use of the computer"[10]. The possibility of homogenizing data by their lemmatization and categorization only came in the 1980s. Dominique Labbée's work [3] implements it in the studies of the vocabulary of the presidents de Gaulle and Mitterrand. Lemmatization, however, is not necessary in all analyzes neither today, it is the purpose of the study (and the cost of the work) that determines it according to [19]. Our study thereafter, for instance, gives an advantage to a non–lemmatized corpus.

Today the most significant work in lexicometry has been produced by Damon Mayaffre [15], [16] etc., who has analyzed comparative recurrences and vocabulary patterns in public speeches of all French presidents from de Gaulle to Chirac [16], and later on up to Emmanuel Macron [11], [17]. In [15], Mayaffre points out, for example, the most frequent words of the presidents that are "over–used" or "under–used" when comparing to other presidents ("*problème*" for Giscard d'Estaing, "*civilisation*" for Pompidou, "*naturellement*" for Chirac, etc.) This kind of statistics concerns of course not only some selected words but a large number of contrasted data. An elaborated cluster analysis at the semantic level gives subsequently birth to the interpretations in terms of socio–psychological and historical profiles of the presidents' discourses; therefore the approach is called "logometric" (e.g. a highly cultural *logos* of Pompidou, a didactical one of Giscard d'Estaing, etc.).

A more canonical "lexicometric" analysis, focusing more on lexical and syntactic relations of the words and texts, in one type of presidential speeches – their New Year speeches – is presented by Jean–Marc Leblanc [4]. Beyond all the details we cannot present, it is noteworthy for us that the author emphasizes the need to consider the textual genre in question, and communicative strategies related to this genre. For instance, the addressing relationship towards the listening public is particularly important in the New Year speech, hence an option to study it by the means of lexicometry is to focus on the differences in use of personal pronouns mentioned by the presidents. The emblematic way of French lexicometry to represent it is an analysis in form of factorial analysis of correspondences that indicates the relative differences in use of pronouns (*tu, toi* ('you') / *je, moi* ('me') / *nous* (we), etc.) between the diverse presidents sub–corpora in see [4] p 148). Moreover, the personal pronouns are an object of study also in the general focus on the analysis of the enunciator's position in the framework of French Discourse Studies [12], and this aspect will not fail to concern our corpus either.

3 Political Tweet as a Research Object

The core of the analyses presented below is always based on the public speeches of the French presidents. We claim, however, that an important part of today's political communication is performed also through social media like Twitter, Facebook, etc. [6], indeed, considers the tweet as a sub–genre of political discourse like an interview or a speech and does not see it just as the "reduction of thought to 140 characters". Thus, during the French presidential elections in 2017 [5] have launched a platform called #Idéo2017 to analyze what the candidates say on social networks with the help of lexicometry tools to identify the lexical fields and dominant themes of the different candidates and to establish thereby "the linguistic profile" of the candidates and what differentiates them. The online platform provided by [5], is of free online access to everyone. It contains four types of analyses throughout 13 selected words in French (*France, state, Republic, people, law, work, freedom, democracy*, etc.). The first analysis offers to identify the use of the chosen word by the different candidates. The results are presented in form of two graphs: one for the computation of specificities (the under–/ over–use of the word by the candidates), and the other one for the frequency of the use of the word by the candidates. The second analysis detects the words associated with the chosen word for all candidates. This analysis of co–occurrences is presented in graphical form of the associated words (see Fig. 1). The third analysis consists of computing the use of the chosen word and its derivatives by the different candidates, and the last analysis creates a word cloud that allows to display the lexicon graphically (cf. a word cloud representing our corpus in Fig 2 thereafter).

Thus, on the #Idéo2017 platform, the use of the word *France*, for instance, by the president we will study thereafter is presented as follows.

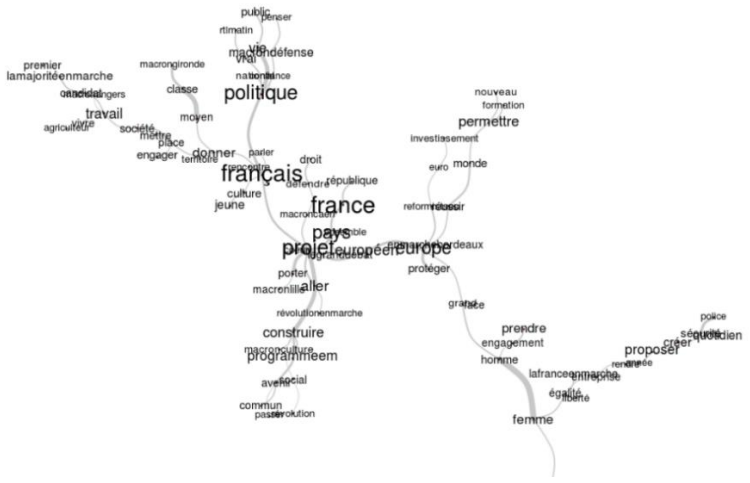

Fig. 1. Graph of the associated words of the word *France* in Emmanuel Macron's tweets in the platform #Idéo2017.

As explained on the platform, the graph represents the relationships between the words in the corpus: the size of the words is proportional to their frequency, and their positioning according to their relationships.

While the position of the words is only indicative to build up the hypothesis, the size of the words is proportional to their frequency also in the graphic display of the vocabulary as word cloud, the representation we will begin with our analysis thereafter (Fig. 3). But, first, we will present our corpus.

4 Our Corpus of #EmmanuelMacron

#Idéo2017 is a platform to study political tweets as samples of a competitive political campaign. Thus, Emmanuel Macron's tweets are only one part of the research on this platform and they are analyzed in comparison to other candidates. Our aim, however, is to study Macron's tweets since he has already become President. Therefore, we have gathered our own corpus of tweets.

The most technical part of the work, first, was getting tweets from Twitter. Since June 2019 Twitter has enforced a limit, therefore only the last 3200 Tweets can be downloaded at a time. For this research, we used the call function get_timelines from rtweet library with 2 parameters: firstly, the twitter username @emmanuelmacron and secondly the maximum number of tweets possible. In order to collect and save the tweets as a xls file for further analysis it is possible to use a script written in R.

To analyze Emmanuel Macron, it is necessary to select the text he has possibly written by himself (or by his PR team) which means excluding all retweets and quotes. It is noteworthy that about 21% of Macron's 3200 tweets were retweets, or partially words of someone else. The remaining tweets still need to be addressed critically since it is unknown whether the text was written by Mr. Macron personally or by his PR team. However, it is the official account of #EmmanuelMacron, hence it can be considered as his discourse.

As for further details, it should be noted that not all text analysis programs, nor the authors of this research do not recognize all the languages or characters used on Twitter. Therefore, all tweets that were not in French, for example, birthday greetings in Arabic or simply URLs, had to be removed. In any case, when Macron expresses himself in a foreign language, we can often find a following tweet expressing the same content in French, thus we consider nothing important is lost. After filtering, there were 2308 of the 3200 tweets left for the analysis. The time frame of the corpus is from December 12, 2017 to December 11, 2019.

5 First Tests and Results

Initially, for the in-depth analysis, Lexico5.8.1 was intended to be used but we decided to use the Voyant Tools in parallel to help us to build up the hypothesis to study. Figure 3 presents a survey of our corpus in the form of a word cloud as follows:

Fig. 2. Word cloud of the tweets by #EmmanuelMacron visualized by Voyant Tools.

First of all, one can see in this visualization a series of quite common words for presidents. *France*, *français*, *Europe*, *république* are listed as highly frequent in speeches of all French presidents [15]. Yet as an accent among these words by Emmanuel Macron comparing to de Gaulle, for example, we can consider maybe the word *monde* ('world'), referring to the general political changes in times. Among more specified words, one to study considering this representation could be the verb *faire* ('to do'), that can be discussed in contrast to the findings by [15] who characterized Chirac as a president of word (*president de dire*), performing in discourse what he lacked in his limitations of the political power. In this contrast, we could say that Macron figures here as a 'president of action'.

However, one of the specific words this graph represents is the verb form *devons* that is the first person plural form of the verb *devoir*, meaning 'have to, should, must etc' or, as a noun, 'duty, obligation or homework'. The verb form *devons* thus means 'we have to'. Linguistically, the non–lemmatized corpus has an advantage here to indicate the over-use of this concrete verbal form: while the first person plural *devons* appears as "over–used", after special checking in data, we can say the verb does not occur in the first singular form at all – no tweet expressing something the President must do himself… The questions about the enunciator position we face here are all the more intriguing in the complex if we consider yet that [11] and [17] have outlined first of all the modal expression *je veux* ('I want') in Macron's public speeches. [8], on the other hand, notes that there are many *nous* ('we') as well in Macron's tweets at least when candidating. Thus, further studies must show how Macron-specific, his Twitter communication-specific or other way specific our findings are.

As in our corpus, we do not have other candidates to compare with, we can study the evolution in time. And, as from the external context we know that an important variable to have possibly an influence on the way the French president can express himself could

be the arise of the Yellow Vests Movement, we decided to study it in more details. To have a moment of juxtaposition, the tweets were split in half: before the Yellow Vests Movement (December 12, 2017 until November 16, 2018) and after the Yellow Vests Movement (November 17, 2018 until December 11, 2019). The first march of the Yellow Vests was on November 17, 2018; therefore, it was decided to split on that date.

Fig. 3. Frequency of the word *devons* before the Yellow Vests Movement represented in Concordance Plot by AntConc.

Fig. 4. Frequency of the word *devons* after the Yellow Vests Movement represented in Concordance Plot by AntConc.

Figure 3 and 4 indicate that the form is rather over-used in first part of the corpus and under–used – or less used – in the second part. Thus, in the socio-discursive perspective we could propose that the Yellow Vests Movement that occurs nearly in the middle of our corpus quite seemingly affects the discourse on Twitter by #EmmanuelMacron.

For a discursive analysis, we can proceed to a contextual study also with AntConc or Voyant Tool by visualizing the textual concordances of the moments we judge important according to the schema, but we must also remember that the indications these programs give are still schematic, they do not give a precise time frame. To create a visual of the total number of tweets, the frequencies of the word *devons* dividing everything into months, by now, Microsoft Office Excel has turned out the easiest and the most accurate tool to use because using it everything can be easily edited on the figure.

Figure 5 represents thus the exact portions of the tweets and the moments to study in the use of *devons.*

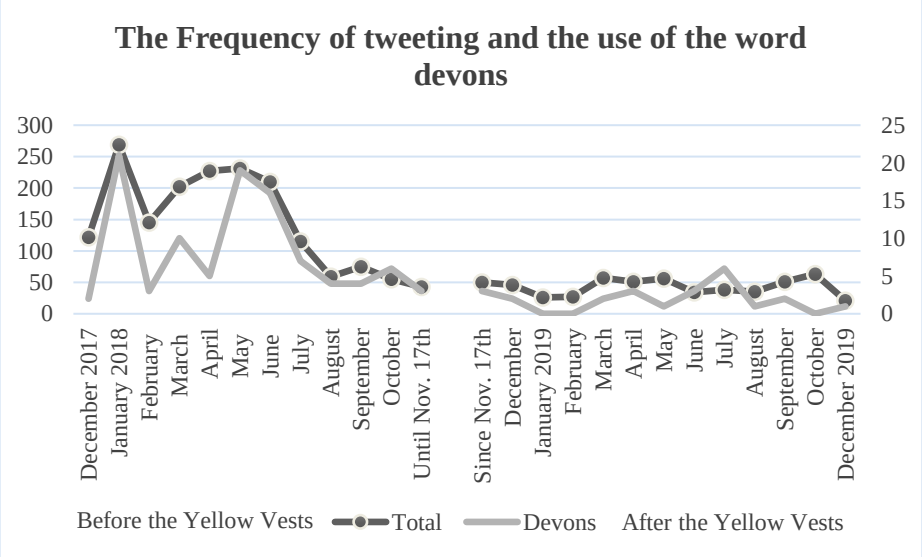

Fig. 5. Maurer 2020. The Frequency of tweeting of the President Macron and the use of the word *devons*.

This figure visualizes the exact points of the abundant use of the word *devons* but also the proportion of the tweets in general that we see have considerably diminished in the second period. Discourse Analysis has thus many aspects to study here.

6 Conclusion

To build up our analysis of the tweets by #EmmanuelMacron, the paper started with an historical overview of the types of analyses conducted in the political lexicometry on the French Discourse Analysis field and in particular as for the French Presidents as research object. Today's most noteworthy works were presented concerning the analysis of the speeches of a series of presidents but also concerning the tweets of the future French president Emmanuel Macron while candidating. We then moved on to the presentation of our own corpus by some simple in-depth analysis tools as word cloud in Voyant Tools and plot concordance in AntConc software. These tools were presented first of all as for their capacity to provide hypothesis but also as for their visual accuracy to represent the data. We can say that Voyant Tools and AntConc were good to propose hypothesis but for more detailed representation, they were inaccurate because they do not give a precise time frame that could be also important to know.

In discussion of the first results, the background of the works conducted in French Discourse Analysis gave a series of paths to consider for the further exploration of the data. We can consider specific vocabulary "over-uses" indicating possibly a broader discusive profile of the president (Macron as the president of *faire* ('to do')). We certainly have to detail the study of tweet as a specific text genre to explain the salient verb

form *devons* (‘we have to’) we discovered in the corpus. Moreover, we can complete the analysis by an enunciation analysis as for other personal pronouns used in the data. Thus, the starting points for the further exploration are promising, and the ways to proceed considered and the solutions found have only motivated the project to keep finding the accurate tools to study the questions these tools have contributed to arise.

Acknowledgements

This work was supported by the Estonian Research Council grant PRG934 “Imagining Crisis Ordinariness: Discourse, Literature, Image”.

References

1. Dacos M., Mounier P.: Humanités numériques. État des lieux et positionnement de la recherche française dans le contexte international, https://www.enssib.fr/bibliotheque–numerique/documents/65357–humanites–numeriques–etat–des–lieux–et–positionnement–de–la–recherche–francaise–dans–le–contexte–international.pdf, last accessed 2020/01/27. Institut français/ministère des Affaires étrangères pour l’action culturelle, Paris (2014).
2. Guilhaumou, J.: Le corpus en analyse de discours: perspective historique. Corpus 1. http://corpus.revues.org/8 (2002).
3. Labée, D.: Le vocabulaire de François Mitterrand. Presses de la Fondation nationale des sciences politiques, Paris (1990).
4. Leblanc, J.-M.: Analyses lexicométriques des vœux présidentiels, ISTE Group, London (2016).
5. Longhi J., Marinica C, Hassine N., Alkhouli A., Borzic B.: The #Idéo2017 platform. In: Proceedings of the 5th conference CMC and Social Media Corpora for the Humanities Bolzano, Italy, 3rd and 4th October 2017, pp. 46–51. halshs–01619236 (2017).
6. Longhi, J.: Essai de caractérisation du tweet politique. L’Information grammaticale 136, 25–32 (2013).
7. Longhi, J.: Humanités, numérique: des corpus au sens, du sens aux corpus. Questions de communication 1(31), 7–17 (2017).
8. Longhi, J.: Le discours d’Emmanuel Macron, construction d’un storytelling. The Conversation 2017/01/26 (2017).
9. Longhi J.: Tweets politiques: corrélation entre forme linguistique et information véhiculée, In: Mercier A. et Pignard-Cheynel N. (eds.), #Info. Partager et commenter l’info sur Twitter et Facebook, pp. 295–314. Editions de la Fondation MSH, Paris (2018).
10. Cotteret, J.-M., Moreau, R.: Recherches sur le vocabulaire du général de Gaulle: Analyse statistique des allocutions radiodiffusées, 1958–1965. Armand Colin, Paris (1969).
11. Lorriaux, A.: Le “je” d'Emmanuel Macron. Interview avec Damon Mayaffre. Sciences Humaines. Août–septembre (2017).
12. Maingueneau, D.: Énonciation et analyse du discours. Corela HS–19, doi: 10.4000/corela.4446 (2016).
13. Maldidier, D.: Analyse linguistique du vocabulaire politique de la guerre d’Algérie d’après six quotidiens parisiens. Thèse de doctorat. http://classiques.uqac.ca/contemporains/maldidier_denise/analyse_linguistique/analyse_linguistique.html, last accessed 2020/01/27. (1969).

14. Marcellesi, J.-B. Éléments pour une analyse contrastive du discours politique. Langages 23, 25–56 (1971).
15. Mayaffre, D.: Paroles de président. Jacques Chirac (1995–2003) et le discours présidentiel sous la Vème République. Champion, Paris (2004).
16. Mayaffre, D.: L'analyse de données textuelles aujourd'hui: du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle. In: Salem, A. ; Fleury, S. (eds) Lexicometrica, pp. 1–12. hal–00551468 (2007).
17. Mayaffre, D. (Interview avec): Emmanuel Macron ne converse pas avec le peuple, il le met à distance, Figaro 2018/12/11 (2018).
18. Pêcheux, M.: Analyse automatique du discours. Dunod, Paris (1969).
19. Salem, A.: Pratique des segments répétés. Essai de statistique textuelle. Klincksieck, Paris (1987).