The 2016 presidential primaries in the United States: a quantitative and qualitative approach to media coverage

Hélène Ledouble, Emmanuel Marty

▶ **To cite this version:**

Hélène Ledouble, Emmanuel Marty. The 2016 presidential primaries in the United States: a quantitative and qualitative approach to media coverage. Studia Neophilologica, 2019. hal-01373150

HAL Id: hal-01373150
https://hal.science/hal-01373150

Submitted on 6 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Preliminary version (April 2016) to be published in STUDIA NEOPHILOLOGICA
https://doi.org/10.1080/00393274.2019.1616219

The 2016 presidential primaries in the United States: a quantitative and qualitative approach to media coverage

Hélène Ledouble and Emmanuel Marty

Abtract

Based on the theories of media frames and their application to discourse analysis, the objective of this study is to identify the discursive strategies of American newspapers in the coverage of the 2016 primary elections. A large corpus of more than 3,000 articles published in 6 national papers between January and March 2016 is analysed using the textual statistical tools Iramuteq (Ratinaud & Dejean 2009) and TXM (Heiden et al. 2010). Several semi-automated operations, based on lexical co-occurrence (notably hierarchical descendant classification and factorial analysis) enable a re-construction of semantic cues (Mayaffre 2014), revealing ideological postures, argumentative strategies and thematic disparities between the different newspapers. In the paper, we address the subjective interpretation schemes inherent to the way each media outlet presents the different candidates depending on their communication contract, with a final focus on the candidate Donald Trump.

Keywords: American presidential election, press coverage, media frames, lexicometrics, lexical cooccurrence, themes, semantics, morphosyntax, Trump

1. Introduction

This article falls within the framework of a research project based on the study of presidential elections in the United States and France as seen by the media. Using the theories of media frames and their application to discourse analysis (Gitlin 1980, Entman 2010), our objective is to characterise editorial strategies through discursive contents in the coverage of primary elections.

In this study, we focus on the 2016 primary elections in the United States, covered by six national papers between January and March 2016. Our corpus of almost three million words is analysed using lexical statistics: several semi-automated operations based on lexical co-occurrence are implemented with a view to revealing ideological postures and thematic differences between different newspapers. We simultaneously address the subjective interpretation schemes used by each media outlet, depending on their editorial identities, as they introduce the different candidates.

Following the presentation of the contextual and theoretical background in the second part of our study, we will delineate the characteristics of our corpus and the methodological approach used in the investigation. The statistical instruments presented in the third section will precede the corpus analysis itself. In the fourth section, an introductory perspective on the different sources composing our corpus will lead us towards a more comprehensive approach of its thematic content.

Version préliminaire

The study finally applies a complementary semantic and morphosyntactic method for a deeper investigation of thematic issues, culminating in a focus on the candidate Donald Trump.

2. Context and research questions

The 2016 presidential elections in the United States received intense media coverage. This coverage started long before the candidate Donald Trump was elected: during the primary period, from January to June 2016, the media began to reveal the agendas and personalities of both the Democrat and Republican candidates.

The role of the media in presidential elections has been an inexhaustible source of controversy for academics worldwide. The question of media influence on people's choices, addressed by Lazarsfeld et al.'s pioneering work (Lazarsfeld et al. 1944), is the core of this controversy. To tackle the issue, we now have strong indicators, inherited from media frame theories (Entman 1993), to determine some of the processes by which the press coverage of an election may highlight certain issues or aspects related to the campaign or to the different candidates.

Our objective in this article is to identify the discursive and lexical content of media frames, as this constitutes the raw material in the perception of election issues, actors and positions. In 'media frame theories' and in their methodological application (frame mapping), discursive content and strategies (even ideologies) are grasped by identifying and measuring lexical co-occurrence. We are therefore interested in lexis as an indicator of discourse strategies (see section 3.2. for methodological explanations). By setting this goal, we also address the following questions: are those frames spread homogeneously across different newspapers, or can we identify differences between them? If so, what kind of political or editorial logic might structure the apparent disparities? And finally, are there specific discursive characteristics in news reports concerning the presidential candidate Donald Trump? To answer these questions, we base our corpus analysis on media frame theories, considered as a heuristic and operational concept for Critical Discourse Analysis (CDA) of media coverage (van Dijk 1993).

Media frames constitute both a theoretical paradigm of social co-construction of reality, in the wake of Goffman's concept of frame (1991), and a heuristic set of methods and tools for media discourse analysis (see section 3.2 for the specific method, and section 4 for its application to our corpus). As far as the theoretical aspects are concerned, the transposition of Goffman's frame analysis to media discourse has been performed by Gitlin, for whom 'Media frames, largely unspoken and unacknowledged, organise the world both for journalists who report it and, to a large degree, for those who rely on their reports' (Gitlin 1980: 8). This principle corresponds to what Gamson and Modigliani (1989: 4) call 'interpretive packages', which 'have the task of constructing meaning over time, incorporating news events into their interpretive frames'.

Since their work, Entman has developed a definition of framing as a process of – necessarily biased – meaning construction, which seems to be widely accepted by academics. According to him,

> [t]o frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described. (Entman 1993: 52)

version préliminaire

Furthermore, a dichotomous definition of frames has been developed by Iyengar (1991: 2), opposing episodic and thematic framing: episodic framing focuses on 'specific events or particular cases' to illustrate an issue, while thematic framing 'places political issues and events in some general context', in order to question them. In the coverage of an election, Iyengar highlights the fact that the media strongly favour episodic frames over thematic ones, by overusing what he calls 'horse race' framing, that is, devoting most of their reports to the succession of polls, alliances and strategies, to the detriment of political and social issues (Iyengar 1991: 133-134).

However, not all newspapers use the same frames and interpretive packages. Indeed, their editorial strategies in the coverage of different issues are also determined by their political values, socio-economic and editorial constraints, all of these being linguistically encoded in media discourse in order to be appropriately decoded by targeted audiences. Consequently, we may observe more or less significant differences in the coverage of primary elections by the media, including their use of frames: linguistic and semantic routines that characterise their discourse.

3. Corpus material and methodology

3.1 *Corpus presentation*

Our corpus comprises 3,151 articles published by six national newspapers in the United States (*The New York Times*, *The Washington Post*, *USA Today*, *New York Daily News*, *The Wall Street Journal*, *New York Post*), from January 25 to March 25, 2016. For that period, we extracted all articles containing the words *primary, primaries, caucus, caucuses*, totalling 2,898,035 words from 3,151 articles.[1] Then, we proceeded to a Descending Hierarchical Classification (DHC, see 3.2.2 for further explanation) on the whole corpus. This exploratory analysis produced results which were lexically very heterogeneous, due to the choice of keywords used for corpus extraction (notably *primary/primaries*). Some articles included lexical material related to other contexts (mostly sports-related), their lexicon being confined to one specific class by the DHC. Text segments constituting this class were consequently removed from the analysis, to focus on the proper context of *primary*, that is, the American presidential election, in a new and more focused sub-corpus. The new sub-corpus totals 1,980,221 words from 3,117 articles. This new corpus can be broken down into the following parts:

Table 1. Corpus information[2]

Media	Word-tokens	Percentage of the whole corpus	Number of articles per newspaper
The New York Times	789,431	39.7%	957
The Washington Post	575,481	29.0%	815
The Wall Street Journal	351,916	17.7%	623
USA Today	110,275	5.5%	234

[1] The international news database Factiva (produced by Dow Jones) was used for this extraction.
[2] The difference in the number of articles and their size is measured (and balanced out) by textual statistics (see section 3.2.1 for more information).

sion préliminaire

New York Post	76,154	3.8%	244
New York Daily News	76,964	3.9%	244

Each of these newspapers focuses on a specific audience and has its own editorial identity. They all relate in different ways to the Conservative or Liberal perspective, depending on this identity and political stance. We aim to investigate this position further in an unbiased manner by studying their divergences or convergences in terms of lexical content, using statistical methods, which we present below.

3.2 *Methods for textual analysis*

Statistical methods are here applied within the CDA theoretical framework. Therefore statistical indexes of word frequency give information on language use in a given communicative situation. As far as the methodological aspects are concerned, two successive operations are central to textual data analysis carried out by any textual statistics application: tokenisation and partition. Tokenisation consists of cutting the text into minimal lexical units, tokens. Identical tokens are then gathered as 'word types' in an index and their occurrences are counted. Partition consists of splitting and gathering texts (here, newspaper articles) according to extra-textual characteristics (in this case, the newspaper that published the article), enabling contrastive analyses based on the different parts of a whole corpus. Thus, textual statistics applications generate a double-entry table which associates lexical forms (rows) with partitioned texts (columns). Most statistical processes are then implemented based on this table, to determine whether the potential imbalances between the different parts are statistically significant (Lebart & Salem 1994).

In this section, we present different approaches to CDA, most using statistical tools – lexical specificity, Correspondence Factor Analysis (CFA) and Descending Hierarchical Classification (DHC) – with a view to revealing how the use of these different methods can assist semantic interpretation of a dense corpus.

Version préliminaire

3.2.1 Lexical specificity and Correspondence Factor Analysis

According to Lebart & Salem (1994) the field of textual statistics is perfectly adapted to discourse analysis and the study of argumentative strategies, ideological postures being potentially conveyed by the repetition of words and phrases. A statistical index named 'lexical specificity index', based on the hypergeometric model (Lafon 1980), identifies significant imbalances in the distribution of words between parts of a whole corpus, by measuring the difference between a theoretical (balanced) distribution of lexical units (as if the extra-textual characteristics had no impact on their position) and their actual distribution.[3] A heuristic way to represent these lexical disparities, then, is the Correspondence Factor Analysis developed by Benzecri (1973). 'Correspondence analysis is a method that gives a geometrical representation of the associations between two sets of elements in correspondence as they appear in a table' (Beaudouin 2016: 8), here a contingency table crossing lexis (in row) and partitions (in columns). 'Correspondence analysis uses a Euclidean space and a distributional distance, or the chi-square distance, which is a distinctive feature of correspondence analysis' (Beaudouin 2016: 9). The CFA graph thus displays lexical distance or proximity between both lexical forms and corpus partitions, spatially.

3.2.2 Descending Hierarchical Classification

The Descending Hierarchical Classification (DHC), developed by Reinert (1983) and implemented in the Iramuteq software (Ratinaud & Dejean 2009) splits the text into segments of equal length (about 40 words)[4] and codifies the presence or absence of each form within each segment so as to gather those containing the same words. The common lexical units that co-occur within segments form different lexical classes, and the significance of their affiliations to those classes are expressed by a χ^2 score. Reinert (2008) calls those classes 'lexical universes': they function as mental and linguistic spaces, closely linked to social representations (Ratinaud & Marchand 2015), within which speakers situate their discourse. This unsupervised approach enables the semantic reconstruction of themes: according to Mayaffre (2008) co-occurential pairs can be considered as 'semantic molecules', a statistically recurring co-presence of forms involving their semantic correlation. Therefore multiple co-occurrences help to disambiguate potentially polysemous words by constructing the previously mentioned lexical universe around them. Identifying the semantic correlation between forms on the basis of lexical co-occurrence constitutes the operational side of media frames analysis, often referred to as 'frame mapping' (Miller 1997).

Additionally, in order to apprehend the linguistic complexity in the expression of emerging themes through textual analysis, we used the TXM platform (Heiden et al. 2010). Its morphosyntactic approach facilitates the identification of the different patterns of information related to a specific element (see 3.3). This syntagmatic and paradigmatic method will prove to be highly relevant for the in-depth study of what we formally characterise as themes (Ben Hamed & Mayaffre 2015).

[3] As opposed to keyness analysis which identifies keywords using a reference corpus, these methods aims at contrasting word distribution among different papers (partitions) taking into account different parameters such as the length (size) of the article and the partition within the global corpus. Specificity will indicate the over-representation or the under-representation of the occurrences of a word (or a sequence of words) in each partition.

[4] This word count is an average, adjusted according to the presence of punctuation marks.

version préliminaire

3.2.3 Similarity analysis

The goal of similarity analysis is to identify co-occurrence between words, providing information on their individual connections within the larger textual structure. Such an approach is conventionally used to describe and visualise social representations (Ratinaud & Marchand, 2015). After calculation of the contingency coefficient, which is a classic similarity index (Flament 1981), the method helps to reveal the organization of textual data, in the form of a 'maximum tree',[5] presenting a fine outline of how to go from one element to the other.

By way of a synthesis, Lexical specificity and Correspondence Factor Analysis are used for the contrastive analysis between newspapers, whereas DHC and similarity analysis are devoted to the process of 'frame mapping' (Matthes & Kohring 2008). Based on this material and these multiple methods for our study, we now turn to the results of our analyses.

4. Corpus analysis

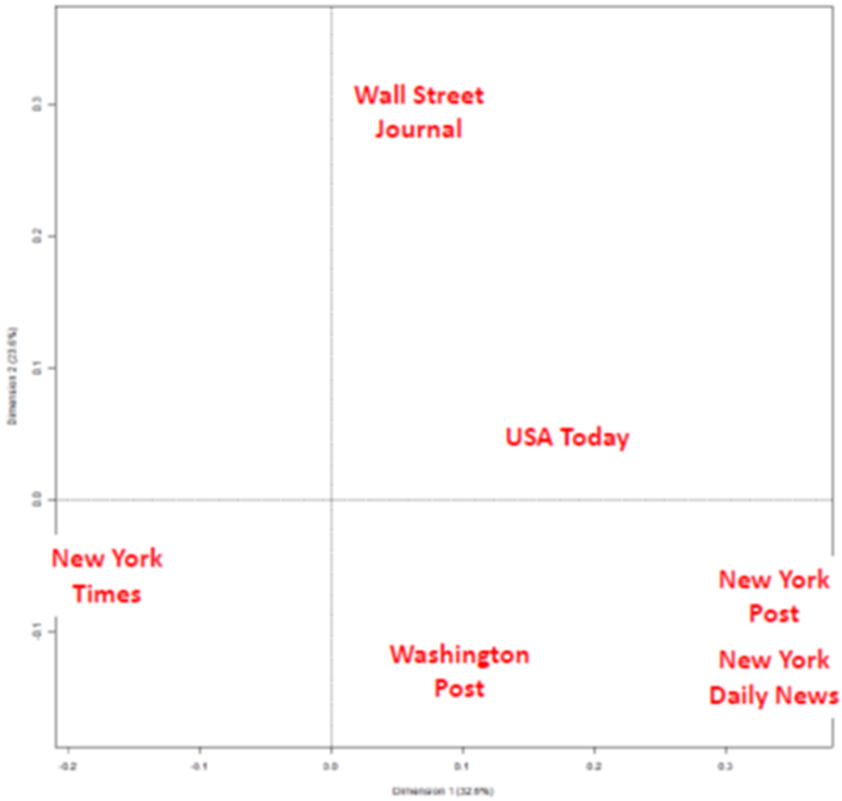
4.1 *General lexical perspective*

Using the statistical index of lexical specificity on this sub-corpus, the Correspondence Factor Analysis graph (Figure 1) represents the lexical proximity (or distance) between sources. Significant imbalances can be observed in the distribution of words between the different newspapers as the graph below shows:

Version préliminaire

[5] Generic term: 'Maximum likelihood phylogenetic tree'

Figure 1. Factorial analysis based on lexical specificities



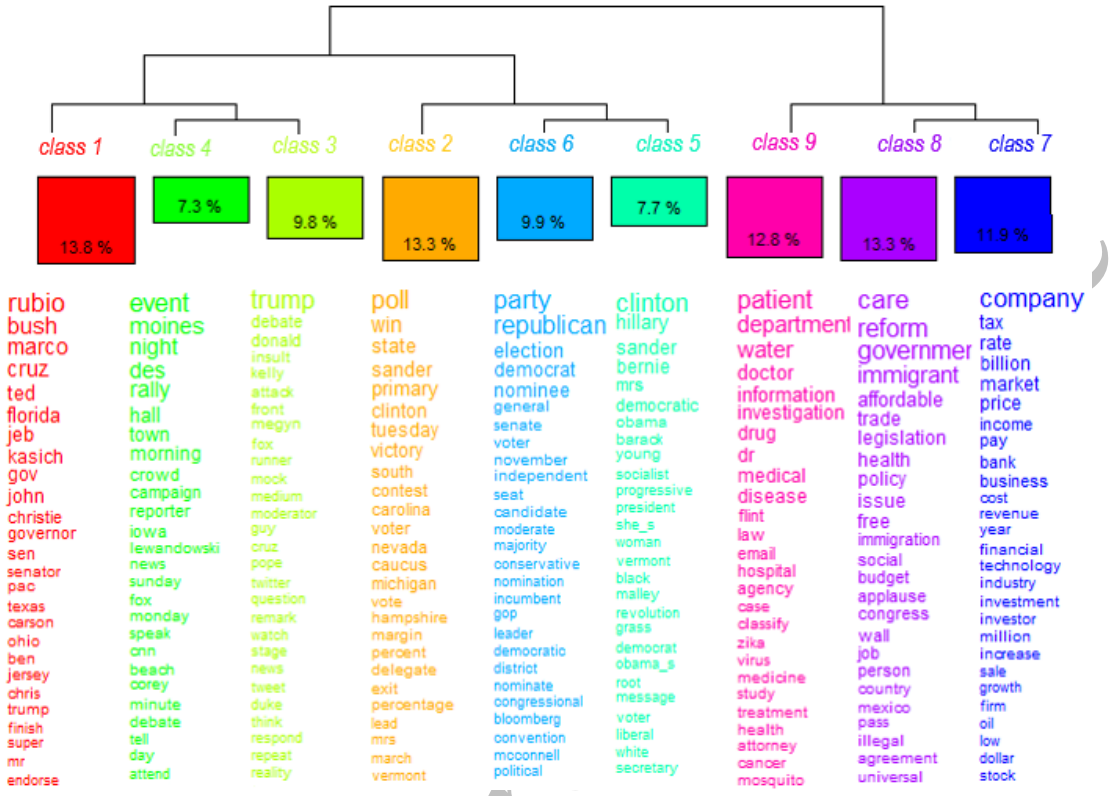
This representation projected on 2 axes is a lexical continuum displaying the different sources and their position (based on word distribution),[6] that requires interpretation: on the horizontal axis, a certain distance can be observed between *The New York Times* (on the extreme left side of the graph), and the *New York Daily News* as well as the *New York Post* (on the far right). On the vertical axis, *The Wall Street Journal* (at the top) is lexically opposed to *The Washington Post* (at the bottom) on the graph.

This can be seen as materialising general differences between the different papers, based on the contribution of each lexical form to a specific source of information. In order to explore this question more deeply and identify specific discursive or lexical components, we now turn to Descending Hierarchical Classification (DHC). Based on lexical co-occurrence, the following class distribution is the outcome of a DHC on our sub-corpus.

Version finale

[6] See section 3.2 for the method brief description, and AFC related literature mentioned in the text for further explanations.

Figure 2. Dendrogram obtained after DHC analysis of the sub-corpus



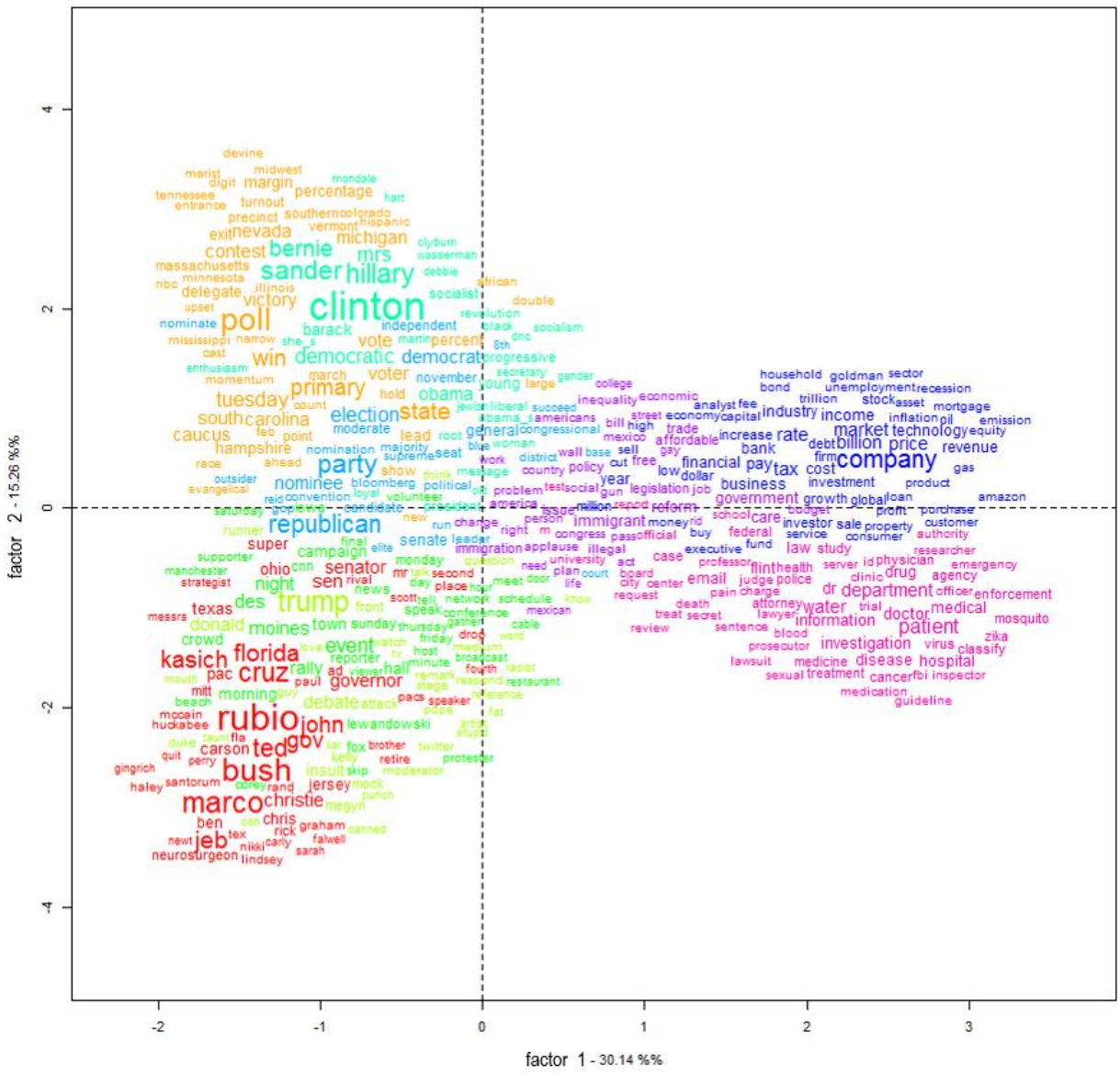
As illustrated in Figure 2, we can distinguish three clusters (three different branches) at the top of the graph, gathering classes presenting a similar lexical content: the first cluster (classes 1, 3, 4), the second cluster (classes 2, 5, 6) and the final one (classes 7, 8, 9). The previously mentioned dissimilarity between episodic and thematic framing is noticeable in this perspective. The first two clusters, on the left hand side of the dendrogram, mostly involve episodic framing, favoring the coverage of polls, candidates' strategies and results before and after each state primary or caucus (*Rubio, Bush, Cruz, rally, attend*, *event,* etc.). This is referred to as a 'horse-race' framing type of information (Iyengar 1991). In the last cluster (on the right), we can identify a very general collection of political and social topics (health, tax, security, etc.) tackled by the different candidates in their programs, corresponding to what we previously designated as thematic framing. But further investigation of these classes is required in order to produce a fine-grained description of the different 'interpretive packages' of these framing models, and a more robust frame mapping analysis.

Version pr

4.1.1 Global frame mapping

Figure 3 (Factorial Analysis) shows the relations between the classes previously identified in Figure 2, and their overlapping sections. We can distinguish the three class-based sections mentioned above:

Figure 3. Factorial analysis obtained as the outcome of DHC upon the sub-corpus

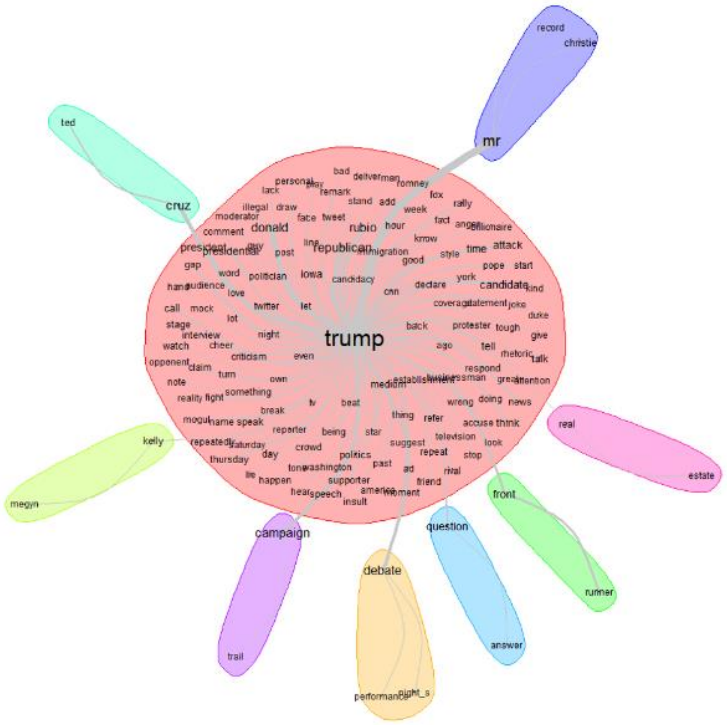


The above representation highlights the proximity between classes (due to their similar lexicon) and the need to focus on their content better to understand the structuration of information in our corpus. We will now focus on each class using similarity analysis, and attempt to define their lexical connections more precisely.

4.1.2 First cluster (classes 1, 4, 3): the Republicans' episodic framing

On the basis of similarity analysis (see section 3.2.3 for methodological explanations), we can identify a specific class (class 3) structured mainly around the Republican candidate Donald Trump – see Figure 4.

Figure 4. Similarity analysis of class 3[7]



The interpretation of this lexical perspective is based on the centrality of a lexical form as well as the width (thickness) of the link between forms (and not their spatial position). Surrounding the central form *Trump* (due to a very high frequency), the related lexicon helps us to understand the semantics of class 3: *Megyn Kelly*, *real estate*, *question/answer*, *Mr Trump*, etc. Trump is presented as a real estate magnate and front runner in the presidential race, and this class is fully devoted to the illustration of his character and actions (with words like *insult, mock, tweet,* etc.), notably in his relations with journalists.

Excerpt (1) exemplifies the lexical universe contained in this class:[8]

(1) Fox News blasted Donald Trump Friday night for his "crude and sexist" tweet trashing anchor Megyn Kelly and for telling his supporters to boycott her program. The network released a statement charging that Trump has a "sick obsession" with the cable news anchor and it's proud of Kelly's work as a journalist. (*New York Post*, 19/03/2016)

[7] The colours are used to separate and highlight the different lexical clusters but do not have a specific meaning

[8] Excerpts are selected on a statistical basis by the textual analysis software: they are representative of lexical classes when they contain a large number of class-specific words.

Version finale

This class is linked to another class, class 4 in the dendrogram (see Figure 2), which focuses on a televised Republican debate held in Des Moines, Iowa. The main feature of the Republican debate was the fact that Donald Trump skipped it, as excerpt (2) explains perfectly:

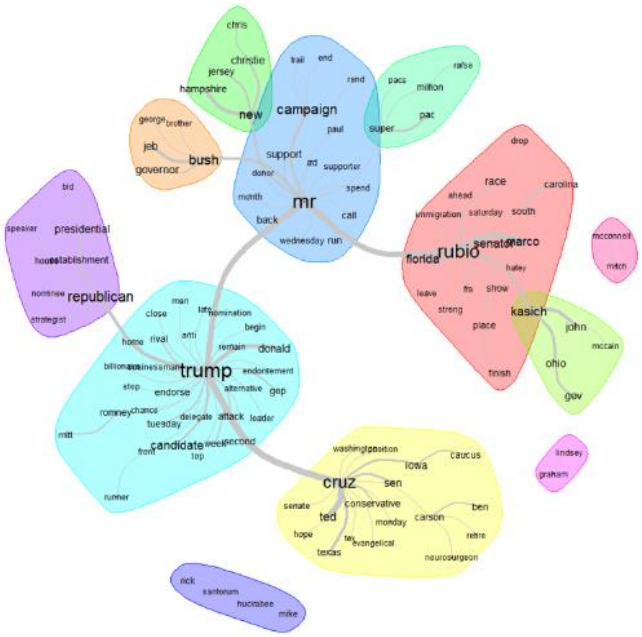
(2) Donald Trump came up early in the Republican presidential debate Thursday, but not often. The seven Republican candidates participating in the Fox News-hosted event from Des Moines quickly addressed the elephant in the room – the absent front-runner, who was boycotting the event in favor of his own rally just 3 miles away. (*New York Daily News*, 29/01/2016)

Representative of the episodic framing, this class is structured around one event, the absence of the main candidate in the debate. This is significant in the sense that much media coverage in the primary period depends upon the political agenda and tackles one-time topics – e.g. his absence in extract (2), a tweet in extract (1) – rather than actively questioning the different political issues. Extract (2) can also account for the importance of the form *Trump* in the class, his repeated name (and wordplay with his name) illustrating his overwhelming media presence:

(3) Trump, Trump, Trump, Trump. Can any conversation of more than 45 seconds' duration fail to turn to le sujet Trump? […] Sunrise tomorrow will be Trump o'clock, the weatherman is calling for a partly Trumpy day, you can read about it in the Trumpington Post on the way home. Wake up and smell the coffee: It's morning in Trumpmerica. (*New York Post*, 06/03/2016)

In this first cluster, class 1 contains all the other candidates' names in the primary, as we can see in the similarity analysis displayed in Figure 5.

Figure 5. Similarity analysis of class 1



Version préliminaire

This is typical of what Iyengar calls 'horse-race framing', reporting daily polls, as the following characteristic segment – extract (4) – illustrates:

(4) [L]atest poll of likely new Hampshire voters (500 Republican and 500 Democratic) in this Tuesday's primary: John Kasich 13%, Jeb Bush 10%, Ted Cruz 7%, Chris Christie 5%, Ben Carson 4%, Carly Fiorina 4%, Bernie Sanders 50%, Marco Rubio 19%, Donald Trump 29% Source: Suffolk University/Boston Globe telephone poll taken Feb. 2-4. (*New York Post*, 07/03/2016)

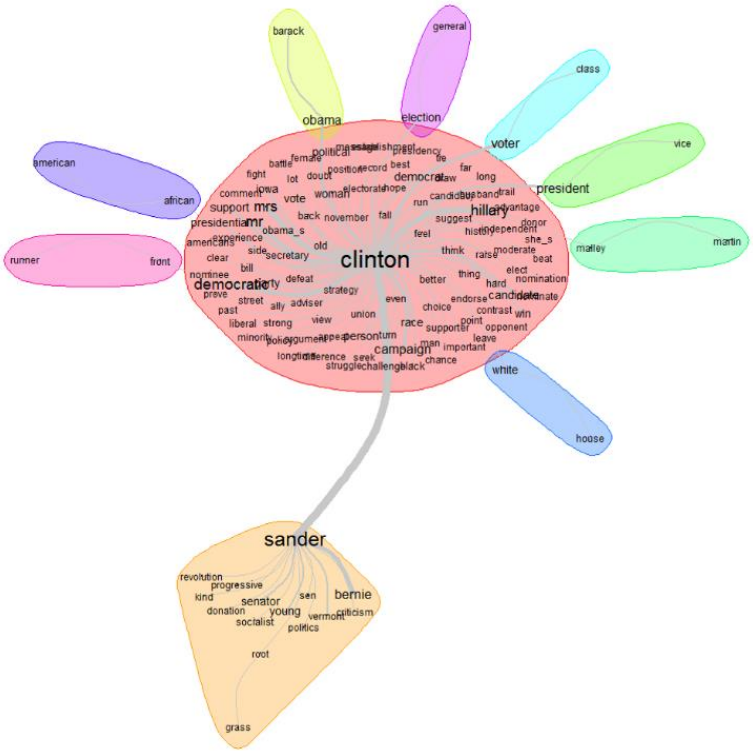
This class is linked to the previous two. Together, they constitute a first cluster of episodic framing. This cluster reveals the predominant focus on the candidate Donald Trump, insofar as he is the structuring element of two classes (including one class linked to his absence during a specific event), and a lesser media coverage of Trump's opponents as all the other candidates end up together in one single class, as opposed to Trump (present in all three classes).

4.1.3 Second cluster (classes 2, 6, 5): the Democrats' episodic framing

The following classes are also part of an episodic cluster, but this time centered more on the Democrats. Class 5 focuses on primary elections among liberal candidates, and contains discourses aiming to define the sociotypes of Democrat supporters (*women*, *black*, *young*, *white*, *liberal*, etc.) and highlighting a duel between Hillary Clinton and Bernie Sanders (the former being more central to the class):

Version préliminaire

Figure 6. Similarity analysis of class 5[9]



Extract (5) is representative of this class as it contains many class-specific words:

(5) What now? Hillary Clinton left the Iowa caucuses with her tail between her legs, besting "democratic socialist" Vermont Sen. Bernie Sanders by just a few votes in the Democratic presidential contest. (*New York Post*, 05/02/2016)

In the same cluster, class 2 is dedicated to polling results and actual caucuses' votes, essentially on the Democrat side (but not exclusively), together with the *Super Tuesday* event, as we can see in excerpt (6):

(6) Hillary Clinton had a tremendous Super Tuesday II, winning in Florida, Illinois, Ohio and North Carolina. Polls had shown tight races in three of the five primary contests, and after a startling loss to Bernie Sanders in Michigan last week, she wasn't expected to win them all. (*The Wall Street Journal*, 17/03/2016)

Class 6 gathers elements that appear somewhat later in the campaign. It is structured around Republicans and Democrats and a prospective (and speculative) analysis of general elections to come, as shown in extract (7):

[9] Note that *Sanders* has been stemmed to *Sander* on this representation (due to general stemming rules).

Version pre

(7) The presidential election moved into a new phase this week, with Democrat Hillary Clinton and Republican Donald Trump emerging as their parties' likely candidates for the general election in November. (*The Washington Post*, 03/03/2016).

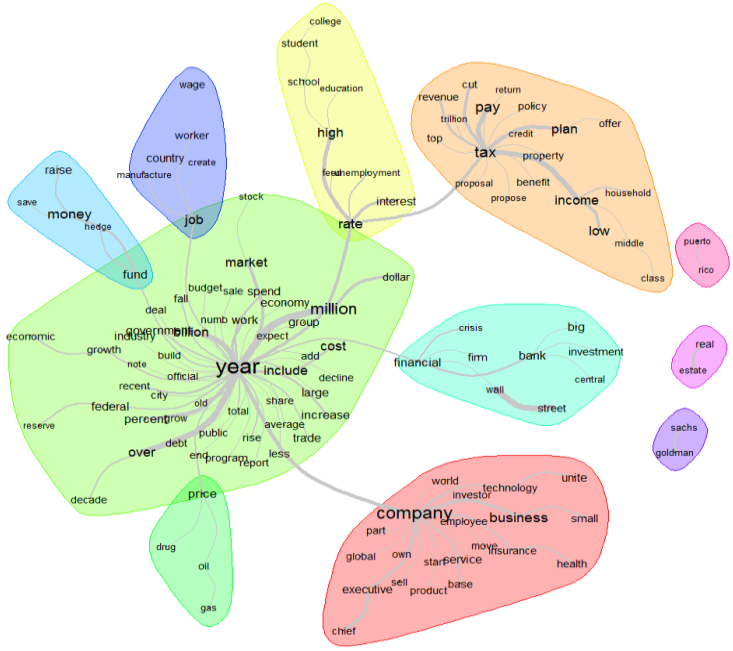
This second cluster, structured by classical episodic framing articles, is made up of sequences that are different from those in the first one, mainly reporting the Clinton/Sanders confrontation in the polls, as the first cluster paid more attention to the early debates, notably among Republican candidates due to the spectacular and often outrageous attitudes of Donald Trump.

4.1.4 Third cluster (classes 9, 8, 7): Thematic framing of political themes

The last cluster embodies what we referred to as thematic framing. Class 7 is structured around the debates on economic, financial and fiscal issues. The economy, with its subsidiary questions of *companies*, *income*, *tax rate*, *gas and oil price*, *banks* and *hedge funds* constitutes the core topic tackled by the different candidates. As we can see in the following figure, the question of time is often associated with these economic issues, *year* being the favoured time measure in this class, as shown by excerpt (8):

(8) To pay for expanded spending on education, energy and family leave, Mrs. Clinton would cap deductions for high-income households, raise taxes on private equity managers' carried interest and increase capital-gains tax rates on assets held between one and six years. (*The Wall Street Journal,* 04/03/2016)

Figure 7. Similarity analysis of class 7



Version préliminaire

Conversely, class 8 focuses on social and societal issues, such as *health care reform* and *health policy*, linked to the question of *immigration*, notably (but not only) from *Mexico*. Excerpt (9) is representative of this type of political content, focusing on programs and proposals for measures:

(9) That's the theory, anyway, and it's deeply embedded in Mr. Sanders's approach. His proposals for single-payer health care, free college tuition and paid family leave financed through a small payroll tax reflect the view that successful programs should be universal and create a connection between individuals and government. (*The New York Times*, 10/02/2016)

The last class (class 9) in this thematic framing cluster includes several events or affairs that seem to go from anecdotal events to real political issues, both being debated or tackled, at the top of which appear Hillary Clinton's email issue, but also local issues such as poisoned tap water from the Flint River in Michigan or the spread of the Zika virus.

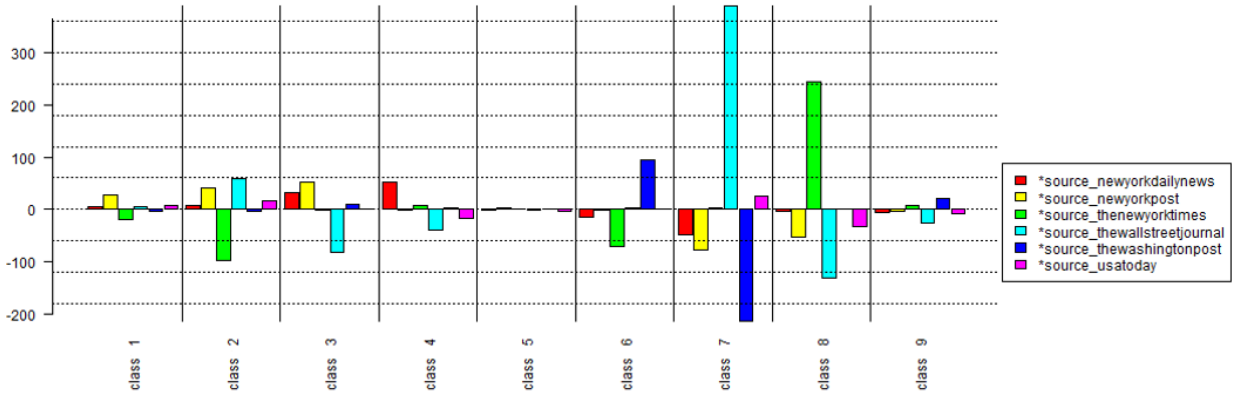
Thematic classes gather elements that show concern about the political content of ongoing debates, rather than sticking to shorter time-frame events, declarations from candidates, voting or polling results.

We will now try to determine if there are significant links between lexical classes or framing types and the editorial identity of the different newspapers, as detailed above (see 4.1), by observing the correlation between each newspaper and the different lexical classes.

4.1.5 Classes and newspapers

Figure 8, based on the χ^2 score, illustrates the correlation between papers and lexical classes:

Figure 8. Correlation between newspapers and lexical classes[10]



The New York Times and *The Wall Street Journal* can be characterised by their predominance in classes 8 and 7 respectively, which are the main two thematic classes (social for one, economic for the other), i.e. placing the primaries within a larger spatiotemporal and conceptual framework.

Conversely, episodic framing seems to characterise the *New York Post* (classes 1, 2 and 3) and the *New York Daily News* (classes 3 and 4).

[10] The size of each bar is determined by the value of the χ^2 score, indicating the over- and under-representativeness of variable modalities (here, the source) in each thematic class (see 3.2 for additional information about the χ^2 score).

Préliminaire

The Washington Post and *USA Today* adopt an intermediate position. Class 6 (episodic from a general perspective) as well as thematic class 9 (dealing with political affairs) both originate from the former. Class 7 on economic debates and class 2 on polling results contain lexicon from the latter.

This result partly echoes some of the positions (i.e. the distance between the *New York Times* and the *New York Post* /the *New York Daily News*) on the first horizontal axis of the Factorial analysis (cf. Figure 1), but also indicates the editorial dynamics by materialising (from a lexical viewpoint) major acknowledged differences between sources of information: on the one hand, *The New York Times*, which, according to Entman, exemplifies a center-left approach to politics, and a source of information committed to the respect of 'objectivity norms' (2010: 390)[11]; whilst on the other hand, the graph displays the distant and opposing position of the *New York Daily News* and the *New York Post* on this horizontal axis. Stylistic factors can probably explain this antagonism, shorter articles and a somewhat sensationalistic approach being specific to the tabloid press.[12]

USA Today can be considered as having a centrist approach to national politics, which would account for its intermediate position on the graph, matching its 'middle-market newspaper' status, as described by the Oxford Index: '[n]ewspapers which are neither upmarket (primarily hard news) nor downmarket (primarily sensationalist), and which combine entertainment with more serious news' (Oxford Index, 2018).

Conversely, the vertical axis of Figure 1 displays the isolated position of the *Wall Street Journal*, in the upper part of this graph, distant and opposed to *The Washington Post*, located in the lower part. This can possibly be explained by their position on the political scale. *The Wall Street Journal* is considered to be more conservative, the target readers often being seen as neocons, while *The Washington Post* is thought to lean towards the liberal side of American politics (together with the *New York Times*). On this same vertical axis, the *New York Daily News* and the *New York Post* would then incline towards a more liberal perspective on American politics, as illustrated by their position at the bottom of the graph. However, their position on the horizontal line – as it is often the case when observing and interpreting CFAs – is the most significant factor here. Their (lexical) proximity to more liberal papers would need additional investigation as well as their respective distinct position on this graph.

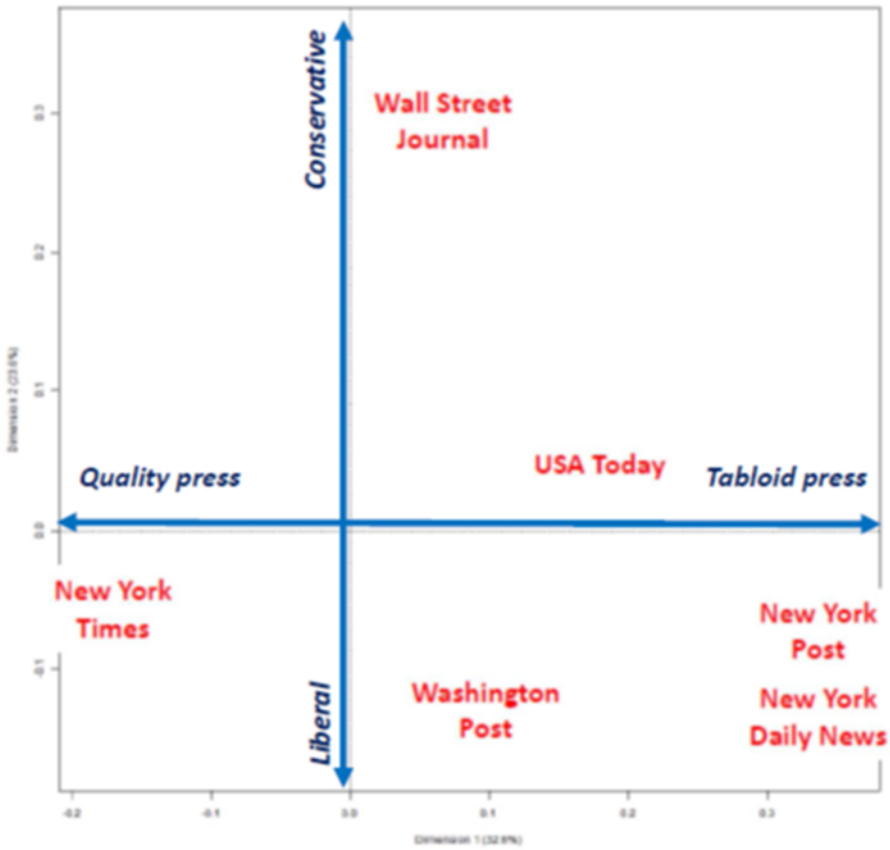
These interpretations could be visualised by naming the continuum illustrated by the two axes as shown in Figure 9:

Figure 9. An interpretation of factorial analysis based on newspapers' lexical specificities

Version préliminaire

[11] He applied the same comment to *The Washington Post*.

[12] 'Newspapers with pages about 30 cm (12 inches) by 40 cm (16 inches), usually characterised by an emphasis on photographs and a concise and often sensational style.' (Collins English Dictionary, 2019)



We can thus observe a correspondence between a content that can be qualified as quality content and lexical universes centered on political issues at stake within the candidates' manifestos.

4.2 *Linguistic approach to themes*

As presented earlier, the DHC enables us to go from words to themes based on lexical co-occurrence. Co-occurring words become indicators of isotopy or semantic correlates, based on a purely inductive – also called non-supervised – perspective. The following method aims at re-defining themes using a meaning-based approach so as to examine their distribution from a contrastive point of view.

4.2.1 Redefining themes

In order to (re)define themes from a more semantic perspective, we used two methodological approaches. We first identified the lexicon in the different classes related to each theme. For instance, for the Immigration theme, we kept the lexical units that relate to this theme in this particular political period (mainly class 8), but which are context-dependent (and not semantically related to Immigration as such): *border, wall, deport, deportation, rapist, undocumented, Mexican*, etc.

Version p

Secondly, to ensure a more semantic approach to themes, we completed the list by adding a limited set of metonyms or semantically-related terms to the main subject: [13] *illegal(ly), immigrant(s), citizenship, migrant(s), visa(s),* etc., thus forming a limited but consistent lexicon encompassing a specific theme.[14]

We applied the same methodology to the following thematic issues: Health, Religion and Gun control (Health: *Social Security, health care, medicare, Obamacare, insurance, Medicaid*, etc. Religion: *religious, religiousness, evangelist, evangelism, evangelical, evangelists, Christian-Baptists, Presbyterian, Pope, pontiff, Christianity*, etc. and Gun control: *gun(s), arm(s), weapon(s), gun control,* etc.)

Figure 10. Correlation between newspapers and themes (based on the specificity index)

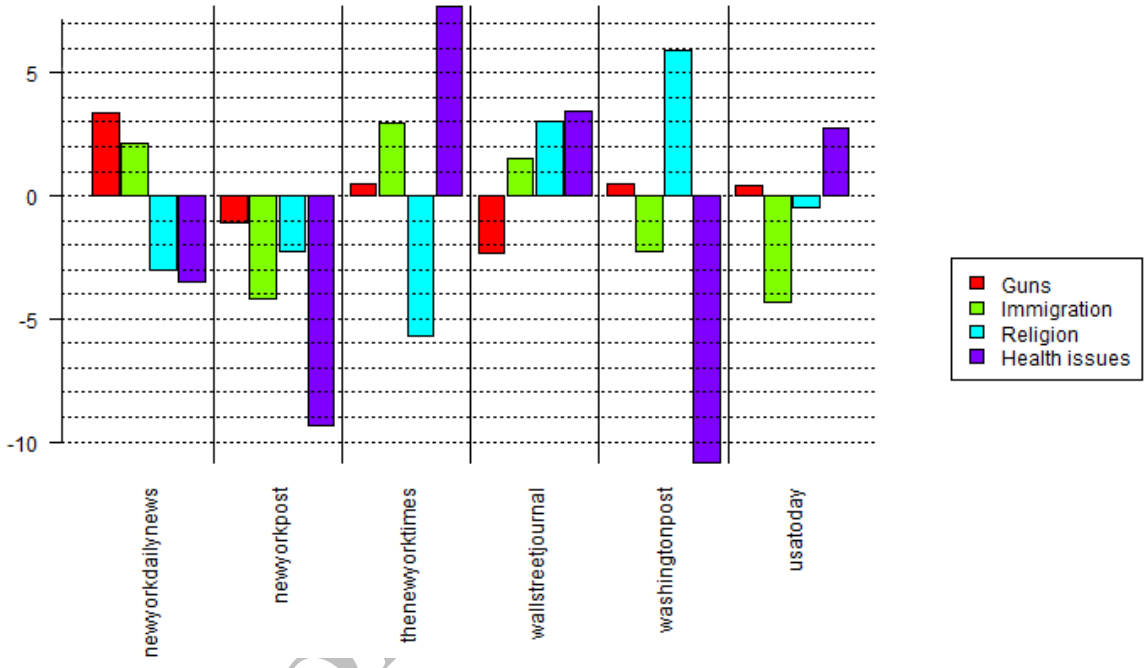


Figure 10 displays the thematic distribution among newspapers. It shows that Immigration and Health issues are particularly dominant in *The Wall Street Journal* and *The New York Times* – see extracts (10) and (11).

(10) Working-class Americans are hostile to free trade and comprehensive immigration reform. (*The New York Times,* 10/02/2016)

(11) His [Sanders] most expensive proposal, about $1.4 trillion annually, is for a "Medicare for all" government-financed health care system, which he said was needed to ensure all Americans have affordable health coverage and to control costs. (*The Wall Street Journal,* 08/02/2016)

[13] We used the Collins and Merriam-Webster thesauri of the English Language for this purpose.

[14] From a general perspective, these units are not always related to the main theme either. However, from a specific point of view (in these 3,117 articles dealing with the US Primary and extracted during that period), the semantic link between these units and the notion of immigration was extremely strong.

Religious issues, as presented in extract (12), are clearly specific to *The Washington Post*'s approach to the campaign:

(12) [B]usinessman Donald Trump […] continued to try to blunt Cruz's appeal among Iowa's powerful evangelical voters (*The Washington Post,* 01/02/2016)

Conversely, none of these issues is statistically significant in the *New York Post*, and only barely addressed in the *New York Daily News,* which is mostly characterised by the handling of gun control issues – see extract (13) – or immigration:

(13) Broadly speaking, the GOP field is wrongheaded in supporting absolutist positions on gun control (*The New York Daily News,* 21/02/2016)

4.2.2 Further investigation of discursive objects

After observing the structuring themes in the different classes, our aim was to further investigate what we qualified as discursive objects. Given the lexical predominance of the Trump theme (see 4.1.2.), we decided to examine what was said about this candidate in two different newspapers and also more generally in all sources.

We used the TXM platform to identify and extract lexical patterns based on part-of-speech sequences: adverbial phrases, complex noun groups, long verbal phrases, etc.). For instance, looking for all the adjectives associated with *Trump* and *style* within 20 words[15] yielded passages such as shown in extract (14):

(14) Trump is belligerent and hyperbolic, with an authoritarian style. (*The Washington Post*, 02/02/2016).

Other complex patterns that can be pointed out include structures such as *Trump* + adverb + verb-future (with potential lexical insertions), exemplified by extract (15):

(15) Donald Trump will very likely be the Republican nominee for president, and there is a non-zero chance he could win in November. (*The Washington Post*, 18/03/2016).

Without quoting all the potential patterns and queries, we now present a synthetic vision of two very different approaches to the candidate *Trump*, exemplified by *The Wall Street Journal* and the *New York Post.*

Trump in *The Wall Street Journal*

In this newspaper, the candidate Trump is mostly referred to with the following qualifiers: *businessman, billionaire, real-estate magnate, estate mogul / tycoon, New York persona*.

The key elements of his speeches are usually reported indirectly, in order not to promote his very provocative language style while still mentioning his favourite issues: *'Trump is light on*

[15] The corresponding CQL (Corpus Query Language) query is:
[word="Trump"][]{1,10}[enpos="JJ"][word="style"]|[enpos="JJ"][word="style"][]{1, 10}[word="Trump"]

Version préliminaire

policy substance, and his supporters have mainly responded to his stick-it-to-the-man style.'; *'The push to the anti-establishment fringe is so strong by Mr. Trump'*.

The candidate is presented as atypical and provocative (*unconventional, freewheeling, blunt, phony, provocative, controversial*, etc.), but the paper remains rather neutral in its presentation of his actions, the candidate being constantly associated with factual but controversial achievements: *Trump Tower, Trump University, Trump magazine, Trump Golf Club, Trump wine,* etc.

Trump in the *New York Post*

The *New York Post* highlights his omnipresence in the media (*Trump told CNN, Trump said, he warned, he replied*, etc.) and provides evidence of his constant social media activity (*Trump tweeted, warned on Twitter, accusatory tweets*, etc.). The candidate's name is the subject and object of criticism (depicted by the media), as the following co-occurring items reveal: *blame, mock, blasted, trashing, caustic barbs*, etc. Similarly, a semantic field associated with his name is clearly related to warfare: *attack, defend, blows, assaults, hostilities, feud, shoot*, etc. Finally, lexical items that can be categorised as insults regularly co-occur with the name Trump, as, in that respect, the initiator of those insults: *egghead, liar, pussy, corrupt, sick, crook, pawn, fat pig*, etc.

As opposed to the other papers, the *New York Post* often quotes his provocative style in direct discourse in order to emphasise the powerful speech acts of this candidate in the presidential race – see extract (16):

(16) "I could stand in the middle of Fifth Avenue and shoot somebody and I wouldn't lose voters". (*The New York Post,* 06/06/2016)

Direct speech also highlights his dialectic approach, based on simplicity and self-evidence: '*we either have a country or we don't. We need a border. We need a wall*.'; *'I think we are weak. We cannot beat ISIS. We should beat ISIS very quickly'*, etc.

5. Conclusion and perspectives

Several conclusions can be drawn from this analysis. If we focus on the object of this study, we have demonstrated the existence of different media frames and links between types of press and types of frames. The main similarity among sources is the hegemonic position of the candidate Trump (even in Democrat universes). The major difference between sources is the way information on this candidate is presented and, above all, the editorial strategies adopted by the different newspapers, leading them to cover the presidential campaign in an episodic or, conversely, thematic framing.

If we now concentrate on the method, three major conclusions summarise the different points:

- General contexts can be visualised (based on specificity and Factorial Analysis), and opposed to local contexts (based on co-occurrences and concordances).
- The study of non-supervised co-occurrences[16] (based on a textual statistics approach) is relevant in the elaboration of a thematic description of a diversified corpus.
- A quantitative approach informs the analysis, a qualitative approach determines the interpretation of texts, constituting a complementary method for meaning construction.

[16] This can also be referred to by the generic term 'Unsupervised hierarchical clustering'.

version préliminaire

Finally, if we consider the object and the method, lexical statistics enable both a fine-grained corpus segmentation (Ratinaud & Marchand 2015), and a dual (macro and micro) perspective on themes.

Version préliminaire

Author's details
Hélène Ledouble, Laboratoire BABEL, Université de Toulon.
1 rue Louis Cotelle, 83200 TOULON
ledouble@univ-tln.fr

Emmanuel Marty, GRESEC, Université Grenoble Alpes.
Institut de la Communication et des Médias, 11 avenue du 8 mai 1945 BP337 38130 Echirolles
emmanuel.marty@univ-grenoble-alpes.fr

REFERENCES

Beaudouin, Valérie. 2016. Statistical analysis of textual data: Benzécri and the French school of data analysis. *Glottometrics* 33, 56–72.

Ben Hamed, Mahé & Damon Mayaffre. 2015. Les thèmes du discours. Du concept à la méthode. *Mots. Les Langages du Politique* 108, 5–13.

Benzécri, Jean-Paul. 1973. *L'Analyse des Données. Tome I. La Taxinomie. Tome II. L'Analyse des Corrispondances.* Paris: Dunod.

Collins English Dictionary. 2019. Definition of 'the tabloid press'. https://www.collinsdictionary.com/dictionary/english/the-tabloid-press. (last accessed on 8 March 2019)

Dijk, Teun A. van. 1993. Principles of critical discourse analysis. *Discourse & Society* 4(2), 249–283.

Entman, Robert M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4), 51–58.

Entman, Robert M. 2010. Media framing biases and political power: Explaining slant in news of campaign 2008. *Journalism* 11(4), 389–408.

Flament, Claude. 1981. L'analyse de similitude, une technique pour les recherches sur les représentations sociales. *Cahiers de Psychologie Cognitive* 1, 375–395.

Gamson, William & Andre Modigliani. 1989. Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology* 95(1), 1–38.

Gitlin, Todd. 1980. *The whole world is watching.* Berkeley, Los Angeles: University of California Press.

Goffman, Erving. 1991. *Les cadres de l'expérience*. Paris: Les Editions De Minuit.

Heiden, Serge, Jean-Philippe Magué & Bénédicte Pincemin. 2010. TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds.), *Proceedings of the 10th International Conference on Statistical Analysis of Textual Data* (JADT) 2(3), 1021–1032. Rome: Edizioni Universitarie di Lettere Economia Diritto.

Iyengar, Shanto. 1991. *Is Anyone Responsible? How Television Frames Political Issues*. Chicago: University of Chicago Press.

Lafon, Pierre. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les Langages du Politique* 1, 127–165.

Lazarsfeld, Paul F., Bernard Berelson & Hazel Gaudet. 1944. *The people's choice: How the voter makes up his mind in a presidential campaign*. New York: Columbia University Press.

Lebart, Ludovic & André Salem. 1994. *Statistique textuelle.* Paris: Dunod.

Version préliminaire

Matthes, Jörg. & Matthias Kohring. 2008. The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication* 58, 258–279.

Mayaffre, Damon. 2008. De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. *Syntaxe & Sémantique* 9, 53–72.

Mayaffre, Damon. 2014. Plaidoyer en faveur de l'analyse de données co(n)textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958–2014). In Inalco-Sorbonne nouvelle, *Proceedings of the 12th International Conference on Statistical Analysis of Textual Data*, 15–32. http://lexicometrica.univ-paris3.fr/jadt/jadt2014/01-ACTES/01-JADT2014.pdf. (last accessed on 22 January 2019).

Miller, Mark M. 1997. Frame mapping and analysis of news coverage of contentious issues. *Social Science Computer Review* 15(4), 367–378.

Oxford Index. 2018. Middle-market newspapers. http://oxfordindex.oup.com/view/10.1093/oi/authority.20110803100156316. (last accessed on 8 March 2019).

Ratinaud, Pierre & Sylvain Dejean. 2009. IRaMuTeQ: Implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. Presented at the *Modélisation Appliquée aux Sciences Humaines et Sociales* (MASHS 2009), Toulouse.

Ratinaud, Pierre & Pascal Marchand. 2015. Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998–2014). *Mots. Les Langages du Politique* 108, 57–77.

Reinert, Max. 1983. Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse des Données* 8(2), 187–198.

Reinert, Max. 2008. Mondes lexicaux stabilisés et analyse statistique de discours. *9es Journées internationales d'Analyse statistique des Données Textuelles* (JADT 2008), 981–993. Lyon: Presses Universitaires de Lyon.

Version préliminaire