

# ANALYSING JOURNALISTIC DISCOURSE AND FINDING OPINIONS SEMI-AUTOMATICALLY?

Fabienne Baidier

► **To cite this version:**

Fabienne Baidier. ANALYSING JOURNALISTIC DISCOURSE AND FINDING OPINIONS SEMI-AUTOMATICALLY?: A CASE STUDY OF THE 2007 AND 2012 PRESIDENTIAL FRENCH CAMPAIGNS. *Journal of Data Mining and Digital Humanities*, Episciences.org, 2014, 2014. <hal-00919370v3>

**HAL Id: hal-00919370**

**<https://hal.archives-ouvertes.fr/hal-00919370v3>**

Submitted on 2 May 2014

**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.

ANALYSING JOURNALISTIC DISCOURSE AND FINDING OPINIONS SEMI-AUTOMATICALLY?  
A CASE STUDY OF THE 2007 AND 2012 FRENCH PRESIDENTIAL CAMPAIGNS

Baider Fabienne  
University of Cyprus  
[fabienne@ucy.ac.cy](mailto:fabienne@ucy.ac.cy)

## Abstract

This research study tested three different NLP technologies to analyze representative journalistic discourse used in the 2007 and 2012 presidential campaigns in France. The analysis focused on the discourse in relation to the candidate's gender and/ or political party. Our findings suggest that using specific software to examine a journalistic corpus can reveal linguistic patterns and choices made on the basis of political affiliation and/or gender stereotypes. These conclusions are drawn from quantitative and qualitative analysis carried out with three different software programs: SEMY, which semi-automatically provides semantic profiles; ANTCONC, which provides useful *Keywords in Context (KWIC)* or abstracts of texts, as well as collocations; TERMOSTAT, which reveals discourse specificities, frequencies and the most common morpho-syntactic patterns. Analysis of our data point to convergent asymmetries between female and male candidates in journalistic discourse (however conditionally) for the 2007 and the 2012 French presidential campaigns. We conclude that social gender (i.e., stereotypical expectations of who will be a typical member of a given category) and / or political favoritism may affect the representation of leadership in discourse, which, in turn, may influence the readership, hence the electorate. Thus the study recommends the use of corpus linguistic tools for the semi-automatic investigation of political texts.

## Introduction

Research analyzing ideology in discourse (Baider & Constantinou 2014; Kohl 2011; Reutenauer *et al.* 2010; Valette 2010; Mussolff 2008) has shown that the indexicality (indirectness) of the construction of prejudices in discourse, i.e., *the lexical and stylistic choices*, can impart an ideological (racist or sexist) semantic prosody to a text without the text being explicitly racist or sexist. Therefore, the *lexical specificities* that characterize a discourse can reveal an ideological position. The purpose of this article is to show how software such as ANTCONC, SEMY and TERMOSTAT (Drouin 2003; Lesage *et al.* 1993) can help to identify the ideological stance of a discourse. Our studies analyzing journalistic data (2007 and 2012 election campaigns in France) show that results found in a semi-automatic way offer a fairly good picture of and basis for comparison of political profiles and campaigns. After a brief review of the literature on interface, language, media and gender issues, the study data and methodology are presented. Our data (journalistic discourse) are drawn from specific candidates participating in the presidential campaign; our methodological tools are capable of identifying lexical preferences (NP extractor, concordance and syntactic parser). In the final section of the article we analyse the results, arguing that such software helps to verify, invalidate or refine results obtained in previous qualitative studies or manual readings: specifically, they offer complementary results to detect specific lexical and stylistic features. Nevertheless, we add the caveat that the

software results and analyses must be considered in conjunction with a close manual reading and a good knowledge of the social context. In this way, we can verify whether the findings are plausible and arrive at the most accurate interpretation of the lexical findings.

## 1. Previous work

According to van Dijk's theory (2006 *et passim*) the overall strategy of an ideological discourse comprises a dichotomous positioning between several entities:

- positive presentation /action of *our* preferred candidate, *our* party, *our* electorate, etc., emphasizing the positive and downplaying the negative;
- negative presentation/ action of the *other* candidate, the *other* party, the *other* electorate, etc., emphasizing the negative and downplaying the positive.

In order to present a specific viewpoint along a certain ideological continuum, different linguistic devices are used, including certain lexical units, grammatical constructions, metaphors, irony, etc. This dichotomous positioning can, in fact, be accomplished in the most indirect fashion.

### 1.1 Indexicality of ideology and linguistic choices

As mentioned in our introduction, ideology in discourse is mainly indexical (van Dijk 2006). Therefore, Vallette and Rastier (2006, 4) were able to identify racist content in online documents by observing language choices that resulted in a text with the illocutionary force of a racist text, but which was not explicitly racist. For instance, with reference to morpho-syntax, indexical racism was characterized by an intensive use of evaluative adverbs and negative adverbs (*nothing, never, much, too much*, etc.), and an intensive use of the first person plural pronoun in order to build an opposition between *us* and *the other(s)*, as well as a predominant usage of pejorative morphemes.

In a similar vein, earlier research on sexism in discourse noted that language choices can indeed build semantic oppositions and discourse asymmetry between the described entities without being obvious (Ochs 1992; Kitzinger 2007; Goodwin 2011). Working on sensationalist discourse (such as in the UK newspaper *The Sun*), Clark (1998) observed the shifting of blame in sexual violence onto the victims by analyzing 'naming' and 'transitivity'<sup>1</sup> (1998, 183-185). She concluded that different connotations of approval and / or legitimacy are expressed by these linguistic patterns. In the field of education, several studies (Lazar 2005; Penelope 1990) have shown that many English-language textbooks use men and boys for the agents or doers, while when they want to depict negative traits they use women and girls. Bergvall (1994) and Macaulay and Brice (1997) have concluded that the examples used by linguists to illustrate syntactical phenomena reveal an under-representation of females especially in examples indicating agency (9%), as well as a negative semantic prosody since they present women as passive, incompetent and talkative. This was despite the recommendations from the LSA (Linguistic Society of America) regarding the construction of examples.

Similarly, in their examination of discourse in scientific texts in French, Michard (1988) and Michard and Ribery (1998) concluded that females were constructed as 'non-agent' entities (1998[2008], 174-184) because of the specific verbs and/or nouns

---

<sup>1</sup> Transitivity is analyzed to decode lack of responsibility or blame; this is carried out by taking into account the process, the participants in the process and the circumstances of the process (see the framework of Halliday 2002).

referring to them, as well as the presence of particular adverbs and adjectives (high incidence of *few, a little, a few, less*). Combining their syntactic and lexical findings, Michard and Ribery concluded that a power relationship between the described entities can be established discursively.

Indeed such an asymmetry may explain the functions and the qualities assigned to one person versus the other in concrete social situations such as the political arena.

### 1.2 *Journalistic discourse, leadership and indexing gender*

Many studies on leadership have concluded that gender plays an important role in the perception of a public *persona*. Some studies argue that, even though the negative perception of female politicians shown by their peers can be attributed to a complex interplay of factors, it can also be partly attributed to the way female politicians draw (if they do) on gendered linguistic resources (Shaw 2013, 39). Meanwhile other studies have found that that among male and female managers who use similar verbal constructions, the women will be viewed rather negatively (Sung 2013; Thimm *et al.* 2003).

Looking at the media, Cotter (2011) studied two key parameters in the news media that would represent a change in the role of women and their visibility in society: bylines (who is entitled to be recognized by name when writing an article) and story topic (what counts as salient in the journalist's world). Both are elements of prestige in the journalistic discourse community. Cotter concluded that despite increased opportunities for women over time, the position occupied by women relative to men, family, and 'the economy has not changed to the degree we sometimes think it has' (2011, 2532). This finding can partially explain the gender disparities observed in journalistic discourse.

Many researchers have noted a discursive asymmetry in relation to female and male politicians (Kahn 1994, 1996; Sreberny-Mohammadi & Rossi 1996; Roncarolo 2000; Coulomb - Gully 2009, 2012; Bertini 2002). Furthermore, most studies that examined French politicians in the written and oral corpus raised the issue of gender--the gender of the entity described, as well as the gender of the entity writing the text. Thus Caillat (2015), Rabatel (2015) and Devriendt (2015) showed how the reported speech in journalistic discourse was adapted differently when comments referred to a female candidate or to a male candidate in the 2012 presidential campaign. Similarly, Barnes and Larrivé (2011), Moores (1997) Baider (2009) examined how female and male politicians are constructed in the French newspapers. They all concluded that discourse practices and lexical choices play a critical role in the construction of the entities described as either incompetent or the opposite, as the perfect leader.

However the ideological distance between the newspaper and the candidate suggests that power struggles are a primary source of asymmetrical treatments as much as gender (Barnes & Larrivé 2011; Baider 2009; Moores 1997): indeed, Barnes and Larrivé (2011, 2507) found that a right-wing newspaper referred to the left-wing female politician (Arlette Laguiller) by her first name alone because of her political affiliation and not because of her gender. Usually, however, such negative treatments are not used unless an important competition is at stake (Moores 1997; Baider 2009).

The above-mentioned studies all have drawn the general conclusion that the indexicality, or indirectness, of ideological positioning renders (semi)-automatic reading difficult. Therefore, how can we evaluate such positioning in discourse? Since

the studies above point to *lexical specificities* in a discourse that (indirectly) represent an ideological position, we could focus our analysis on the most frequent and specific lexical units in our data to test the hypothesis that these units will represent a certain point of view.

### 1.3 *Indexicality and corpus linguistics*

Much research on political analysis has used software to extract lexical units with the aim of identifying patterns, which might then, in turn, explain either the motivation for the word choice or the impact of the word choice. They all conclude that rhetoric and word choices are fundamental to supporting political strategies. For instance, the typical style of political journalists, which reports on political life primarily through quotations or small excerpts, leaving aside much of the rest of the speech confirms this finding (Saitta 2008). Brasart (1998), in his study of ‘petites phrases’ in French political discourse, demonstrated the existence of linguistic practices used by politicians to make an impression on the audience. Semi-automatic readings allow us to examine how and to what degree politicians fit these expressions into their strategic rhetoric; they know that their words will often be reduced to a few sentences and they might well anticipate this and prepare for it (Krieg Planke 2011).

Quantitative observation can help us evaluate changes in political speeches, allowing us to bypass the axiological dimension that frequently marks both the discourse and its analysis. Corpus analysis using semi-automatic reading can also assess the evolution of the professionalization of communicative labor policies and underlying standards expressible in politics. For instance, Marchand (2009) showed that computer-assisted discourse analysis (statistical text analysis, morphosyntactic analysis) can contribute to the diagnosis of genre, style and attitude in language. A recent study of statements made by the French prime ministers of the Fifth Republic (Dupuy & Marchand 2013) showed that by looking at frequency and specificity they were able to note a shift in lexical choices. Their study concluded that this change in word choices would also reflect a change in the focus of the presidential debates in France: and, in fact, since 2000 debates have moved away from issues closely related to the social fabric such as education and environment, to focusing on economic and financial issues (even when referring to education). This focus is one more commonly found among managers and economists, which could be interpreted as a move towards neo-liberalism on the part of mainstream politicians.

Interpretation of such data requires interdisciplinary knowledge, and this has been true since the first major work published in the field, by Kucera and Francis (1967), who subjected their data to a variety of computational analyses, combining elements of linguistics, language teaching, psychology, statistics, and sociology. This interdisciplinary analysis can be carried out within a framework that integrates findings at the micro-level of language into the macro-level of society (societal rules, social actors involved, etc.). Therefore in our analysis of the data we will consider language as a form of social practice and will apply principles of critical discourse analysis (CDA) (Fairclough 1989, 22), i.e., we will consider that linguistic data are not external to society but are, in fact the opposite, conditioned by other (non-linguistic) parts of society.

## 2. **Corpus, Methodology and Tools**

### 2.1 *Data and data analysis*

CDA explores the often ‘opaque relationships of causality and determination between discursive practices and wider social and cultural structures, relations and processes’ (Fairclough 1995, 132; Widdowson 2000). Such practices arise out of and are ideologically shaped by relationships and struggles related to power (Fairclough 1995, 133), while the indexicality of such relationships between discourse and society would itself be a factor securing hegemony (Fairclough 1995, *ibid.*; Ochs 1992). Our analysis will consider the macro-context in which the discourse occurs, i.e., its social reality, after having studied the micro-level, i.e., the text itself, collocations, frequencies, etc. Therefore we will interpret the data taking into account the social reality of French politics, the French political culture and journalistic practices.

The purpose of the next two sections is to present our data and show how software such as SEMY (Grzesitchak *et al.* 2007)<sup>2</sup> or TERMOSTAT (Drouin 2003; Lesage *et al.* 1993) can help identify either the recurrent semantic elements or the specificities of the data and, as a result, point to the ideological positioning of discourses in journalistic corpora.

We undertook three separate studies focused on the journalistic discourse during the 2007 and 2012 presidential campaigns in France, campaigns where female candidates had a chance to win-- whether in the primaries of their party (Aubry in 2012) or in the second round (Royal in 2007).

The first study contrasts the contextual environment of the names of the three main candidates during the 2007 presidential campaign (Bayrou, Royal, Sarkozy) in order to assess their semantic profile in the journalistic discourse, and to assess whether gender was a factor in the word choices by journalists.

The second study follows the same pattern but focuses on the textual environment of the candidates in the smaller parties (i.e., Communist, National Front, Green, etc.), to see if the same patterns occur as those identified in the first study.

The third study contrasts the contextual environment of the names of the two main candidates in the 2012 socialist primaries (Aubry, Hollande) to check whether two candidates from the same party would be presented with different discourse strategies in the same newspapers.

Our data for the 2007 presidential campaign (partially published in Baider 2009, and Baider & Jacquy 2010) were collected over the 12 months prior to that year’s elections; data was collected for the candidates Royal, Bayrou and Sarkozy as recorded in the newspaper *Le Monde*. We collected 330 articles (roughly 100 for each candidate) that focused on a candidate, i.e., containing the name of the candidate in the title and several times in the text. All candidates were represented by a comparable word count (roughly 19, 000 words for each candidate, almost 60, 000 words). This methodology and size of data are compatible and comparable to the earlier studies in language, gender and politics referred to earlier (section 1.2). These studies also used relatively small corpora (Barnes & Larrivé worked on roughly 32,000 words; Fracchiola focused on one debate since her aim was a qualitative study of politeness strategies).<sup>3</sup> Moreover, Barnes & Larrivé used the same methodology as we did (i.e., they started out with the name of the politician in the title and then looked at the focus

---

<sup>2</sup> Developed at the ATILF at <http://www.atilf.fr/>

<sup>3</sup> Cotter (2011) does not specify the size of her corpus but mentions a ‘small corpus of stories’ (2011, 2520).

of the article with the repetition of the name in the body of the text), making our methodology and corpora directly comparable.

Our data for the 2012 socialist party primaries was taken from the *Factiva* database: twenty (20) articles for each candidate from each of six different newspapers (national dailies: *Libération*, *Le Monde*, *Le Figaro*, *La Croix*; regional dailies *Ouest-France*, *Est-Républicain*). Articles focusing on Aubry and Hollande were chosen from periods considered as the most decisive for the socialist primaries, i.e., March-May 2011 is the period before the Strauss-Kahn fiasco<sup>4</sup>; June-July 2011 is the period just after the DSK announcement that he would not run; September-October 2011 is the time of the socialist primaries. There are two versions of every folder on the politicians: the first version contains all the articles for each candidate sorted according to newspaper; the second folder is text alone, with no details on the source. We have a total of 120 articles for Aubry and Hollande.

Previous research has suggested that analyzing discourse practices through a focus on particular on noun phrases (NPs) referring to female and male politicians could be a way to assess the ‘linguistic profile’ of politicians in journalistic discourse, especially their agency (Eckert & McConnell-Ginet 2003). Agency is fundamental in discourse, since one could assume that to be elected, the candidate must be perceived and constructed as a leader (associations with dynamism, initiative, courage, etc.), which in effect means to be constructed and perceived as an agent, someone in control of the environment. This semantic value also has a pragmatic outcome: the described entity will more likely be perceived as a leader and therefore as the candidate to be elected. In order to assess the lexical environment of each NP under study, we used tools that can process a considerable amount of data, thus enabling us to quickly identify lexical choices. This makes a quantitative and a qualitative reading very feasible.

## 2.2 Methodology and Tools

Frequency and specificity are the most common criteria for corpus analysis, as previous studies on ideology and discourse have noted (Marchand & Ratinaud 2012; Michard & Ribery 1998[2008]; Valette & Rastier 2006). Moreover, lexical units such as derogative adverbs and negative adjectives are considered the most relevant to ideological loading (Reutenauer *et al.* 2010; Valette 2010) and more likely to be found in references to a ‘subordinate entity.’ Regarding methodology, most recent research using corpus linguistic methodology for study of the 2012 presidential campaign (e.g., Marchand & Ratinaud 2012) suggests that discursive strategies can be understood through an investigation based on semi-automatic reading.

Indeed, in recent years several software packages have been produced that are able to extract lexical and syntactic patterns characteristics of the type of data under investigation in the present study. Gender specificities were identified by using these programs, specifically, ANTCO for providing statistical concordance functions; TERMOSTAT for term extraction; and SEMY for extraction of semantic fields. These programs were chosen because they were among the most readily available and free

---

<sup>4</sup> Dominique Strauss-Kahn (known also as DSK), the likely winner of the presidential elections for the socialist party, was charged with sexual assault in May 2011 and therefore could not run in the socialist primaries.

of charge.<sup>5</sup> Further, these are also the programs with which the author is most familiar.

*ANTCONC* : *ANTCONC*<sup>6</sup> provides concordances, collocations, frequency lists, word lists and key word in context (i.e., the line in which the word was used). Especially when there are many concordances, *ANTCONC* is user-friendly and quickly downloaded from the Internet onto the computer. Most importantly for researchers working on different alphabets, the software supports many different languages, including those using the Cyrillic (Greek, Russian) as well as Chinese, Japanese, Korean, etc., alphabets.

*ANTCONC* enabled us to carry out a qualitative analysis since we can isolate blocks of text where the NPs are present; in this way we can visualize the contexts of use of any key word (see Graph 1 below, which illustrates what is defined as a key word in context or KWIC). A manual and selective reading is then possible.



Graph 1 : KWIC abstract for Aubry's name in *le Monde*

The software also extracts and displays collocations, frequencies and word lists. In particular, the categories 'collocates' and 'cluster' allowed us to check whether some NP are more often associated than others with evaluative lexical units (see Michard & Ribéry 1987; Valette 2010). We can very quickly verify a higher number of adverbs, adjectives and negative modifiers such as *un peu*, *quelques*, *moins* collocated with the NPs under investigation. Moreover, a manual reading can point out the *syntactic processes* most associated with these NPs (subject of a sentence for instance) and the agency of a unit (among other thematic roles, cf. Dowty 1991). Therefore, with these abstracts we can observe whether NPs are constructed in discourse as a non-animated entity or at least as less in control of the verb (if they are most often subjects of a verb but non -agents, for instance).

*TERMOSTAT*: *TERMOSTAT* (Lesage *et al.* 1993)<sup>7</sup> identifies grammatical categories and allows us to quickly see which nouns, adjectives or adverbs are the most used or the most specific in the data. Although the software presents features complementary to *ANTCONC*, it does not fully support languages other than French and English. The *TERMOSTAT* platform extracts the frequency and specificity of simple and complex lexical units relative to a reference corpus. The data integrated to the *TERMOSTAT* platform as reference corpus includes 28 million occurrences from *Le Monde* (2002)

<sup>5</sup> There are many tools used in corpus linguistics. Among the free engines, there are *Iramuteq* (University of Toulouse) <http://iramuteq.org/> for French, and *Lexico 3* (University of Paris3) <http://www.tal.univ-paris3.fr/lexico/> for French and English, as well as TXM Unicode-XML-TEI text/corpus analysis platform which the most complete tool for the Humanities among the freeware: <http://sourceforge.net/projects/txm/files/corpora/brown/>. Among the tools for corpus linguistics which are not free, Sketch engine is (according to the author of this paper) is efficient and user-friendly; it also supports most languages: <http://www.sketchengine.co.uk/>

<sup>6</sup> <http://www.antlab.sci.waseda.ac.jp/software.html>

<sup>7</sup> [http://olst.ling.umontreal.ca/~drouinp/termostat\\_web\\_or](http://olst.ling.umontreal.ca/~drouinp/termostat_web_or)  
[http://termostat.ling.umontreal.ca/doc\\_termostat\\_en.html](http://termostat.ling.umontreal.ca/doc_termostat_en.html).



(Drouin 2003)<sup>8</sup>. From this we can identify a list of the most salient terms in our data, either because they are the most frequent or because they are the most specific to the discourse. TERMOSTAT was developed from the research of Lebart and Salem (1994), who quantified the specificity of a term based on its *test-value*, which is a standardized view of the frequency of the lexical units (Drouin 2003, 3). We used TERMOSTAT to examine the most specific lexical categories for each NP investigated (i.e., Aubry, Hollande), criteria which are the basis of our hypothesis. We can subsequently determine which categories are most relevant to a qualitative reading. In Table 1 below we can observe the most specific lexical units for the name Hollande, when compared with the reference corpus of TERMOSTAT:

| Candidate                | Frequency | Specificity | Spellings                 | Morphosyntax categ. |
|--------------------------|-----------|-------------|---------------------------|---------------------|
| primaire                 | 248       | 215.29      | primaire__primaires       | N                   |
| primaire socialiste      | 68        | 159.96      | primaire socialiste       | NAdj                |
| strauss-kahniens         | 35        | 113.96      | strauss-kahniens          | N                   |
| candidature              | 153       | 67.28       | candidature__candidatures | N                   |
| président normal         | 10        | 58.7        | président normal          | NAdj                |
| favori                   | 55        | 56.29       | favori__favoris           | N                   |
| ex-premier secrétaire    | 9         | 52.5        | ex-premier secrétaire     | NN                  |
| de hollande              | 8         | 51.81       | de hollande               | N                   |
| candidat                 | 290       | 48.34       | candidat__candidate       | N                   |
| candidat normal          | 7         | 48          | candidat normal           | NAdj                |
| projet socialiste        | 15        | 44.6        | projet socialiste         | NAdj                |
| investiture socialiste   | 8         | 44.14       | investiture socialiste    | NAdj                |
| présidence normale       | 6         | 43.87       | présidence normale        | NAdj                |
| sympathisants            | 42        | 41.83       | sympathisants             | N                   |
| strauss-kahniens         | 6         | 40.6        | strauss-kahniens          | N                   |
| secrétaire               | 170       | 39.66       | secrétaire__secrétaires   | N                   |
| candidat déclaré         | 13        | 39.42       | candidat déclaré          | NPPAdj              |
| présidentiable           | 13        | 39.42       | présidentiable            | N                   |
| humour corrézien         | 5         | 39.32       | humour corrézien          | NAdj                |
| député                   | 150       | 39.11       | député__députés           | N                   |
| sondage                  | 97        | 37.69       | sondage__sondages         | N                   |
| candidat de substitution | 4         | 34.19       | candidat de substitution  | N                   |

**Table1**  
Result with TERMOSTAT: NP *Hollande* in March-July 2011, in six newspapers

While linguistic expressions such as *primaire socialiste* ‘socialist primaries’ and *investiture* ‘nomination, investiture’ are expressions that we might expect to be specificities of a political text, in order to understand expressions such as *président normal* and *humour corrézien*, we must have some knowledge of the candidate profile. Indeed both are topoi that specifically characterize Hollande, whose campaign was organized around the axis of ‘normality,’ and whose humor, as well as the fact that he was responsible for the Corréze district, were well known by the French. Nevertheless, even if one knows nothing of the candidate, the incongruity of both linguistic expressions and the fact that they are specificities of the candidate profile will trigger the attention of the researcher.

As we can see in Table 2, there is also other information relating to the linguistic profile of the data provided by TERMOSTAT such as<sup>9</sup>:

- the various ‘physical appearances’ of this expression in the corpus (plural or feminine forms for nouns for instance);

<sup>8</sup> See Lebart & Salem (1994) for explanation on the two criteria used to quantify the specificity of a term, one of them being the “standardized view of the frequency of the lexical units”.

<sup>9</sup> Other features include the most frequent compound; the most frequent morpho-syntactic pattern; etc.  
[http://termostat.ling.umontreal.ca/doc\\_termostat/doc\\_termostat\\_en.html#inscription](http://termostat.ling.umontreal.ca/doc_termostat/doc_termostat_en.html#inscription)

- the grammatical category of the lexical unit.

These features allow a thorough quantitative reading of a corpus and enable comparison between corpora.

*SEMY*: *SEMY* is one of the rare software programs that is semantic based and is applicable to the French language. It is unfortunately only available at the ATILF. Working with semantic fields is important and *SEMY* aims to give a semi-automatic semantic profile of texts or linguistic entities. Indeed when fed a text or a noun, the platform automatically extracts a lexicon based on the digitalized *Trésor de la langue française* and gives the semantic field of each word in the shape of a list which Rastier would call a SEMEME.<sup>10</sup> The list provided is very similar to the one provided by TERMOSTAT, although it describes semic features and not lexical units. On the basis of the textual files provided, *SEMY* will give lists based on the semantic recurrence in each file, i.e., using the most salient semantic feature in the data. The researcher is then tasked with making sense of all the units presented; he/she must find a correlation in order to arrive at a semantic profile. To do so, we group together the semic features found in the list using the CRISCO synonym dictionary.<sup>11</sup> For instance, the verb *communiquer* ‘to communicate’ was found as a recurrent semic feature in Royal’s *SEMY*. We look into CRISCO and find a series of verbs which are in the same semantic field as *communiquer*, such as *montrer* ‘to show’ or *déclarer* ‘to declare.’ In turn some of these verbs are also found in *SEMY*’S list for Royal, and we can then organize the data accordingly, in the same semantic field. These three verbs were very much present in the Royal data, and much less so in the Sarkozy data. This grouping, we suggest, allows us to describe the semantic profile of the NPs Royal and Sarkozy in the journalistic discourse under study.

|                   |  |
|-------------------|--|
| <u>montrer</u>    |  |
| <u>indiquer</u>   |  |
| <u>exposer</u>    |  |
| <u>signaler</u>   |  |
| <u>dire</u>       |  |
| <u>afficher</u>   |  |
| <u>déclarer</u>   |  |
| <u>découvrir</u>  |  |
| <u>annoncer</u>   |  |
| <u>manifester</u> |  |

Table 2 Most frequent synonyms for the verb *communiquer* ‘to communicate’

In the next section we report the data found with each software program and discuss our results.

### 3. Report and discussion of the results

#### 3.1 *Does the semantic profiling of a noun indicate gender stereotyping?*

<sup>10</sup> <http://www.atilf.fr/spip.php?article894>. *SEMY* was the result of a doctoral thesis and its patent is pending. For further details please contact [evelyne.jacquey@atilf.fr](mailto:evelyne.jacquey@atilf.fr)

<sup>11</sup> <http://www.crisco.unicaen.fr/des/>

We used both SEMY and the 2007 data to work on the semantic profile of an entity in a text since the stakes were very high: Royal was the first woman candidate to make it to the first presidential round, and she was competing against Sarkozy.

During the campaign Royal complained about unfair treatment by the media, saying that they, for instance, portrayed her failures or gaffes much more often than Sarkozy's. And, in fact, the newspaper *Le Monde* apologised for focusing on her 'blunders' (Baider 2009).

As mentioned earlier (section 2.3, SEMY), the most salient dimension of Royal's candidate personality was the relational aspect. Therefore we used the CRISCO dictionary to categorize the semic features, which then allowed identification of the semantic fields for the candidates according to SEMY (Table 3)

| CANDIDATE NAMES | SEMANTIC FIELDS           | RECURRENT SEMIC FEATURES IN CORPUS  |
|-----------------|---------------------------|---|
| Nicolas Sarkozy | <b>Planning</b>           | <i>viser</i> "to aim", <i>chercher</i> "to seek", <i>poser</i> "to set up", <i>vouloir</i> "to want", <i>envisager</i> "to envision", <i>compter</i> "to reckon"                  |
|                 | <b>Argumentation</b>      | <i>Entretenir</i> "to converse", <i>contredire</i> "to contradict", <i>jouer</i> "to play (on words)"   |
|                 | <b>Action</b>             | <i>signer</i> "to sign", <i>embaucher</i> "to hire", <i>agir</i> "to act" <i>agir</i> , <i>embaucher</i> "to hire", <i>partir</i> "to leave"                                      |
| Ségolène Royal  | <b>Political campaign</b> | <i>Voter</i> "to vote", <i>préférer</i> "to prefer", <i>élire</i> "to elect", <i>désigner</i> "to appoint",   |
|                 | <b>Communication</b>      | <i>Appeler</i> "to call", <i>déclarer</i> "to declare", <i>annoncer</i> "to announce", <i>exprimer</i> "to express", <i>proposer</i> "to suggest", <i>téléviser</i> "to televise" |
|                 | <b>Defensive</b>          | <i>Inquiéter</i> "to worry", <i>renoncer</i> "to giveup", <i>défendre</i> "to defend"   |
| François Bayrou | <b>Debate</b>             | <i>estimer</i> "to evaluate", <i>être favorable</i> "to be in favour", <i>être contre</i> "to be against", <i>denoncer</i> "to denounce"  |
|                 | <b>Progress</b>           | <i>progresser</i> "to progress", <i>conforter</i> "to consolidate", <i>poursuivre</i> "to pursue", <i>proposer</i> "to suggest" and <i>gagner</i> "to win")                       |

Table 3. Semantic fields identified with SEMY (Baider & Jacquey 2010)

In the context of the name Royal, 50% of the features listed by SEMY belonged either to the field of COMMUNICATION (such as *to call*, *to declare*, *to announce*, *to express*, *to suggest*), or to the field of POLITICAL CAMPAIGN (*to vote*, *to prefer*, *to elect* and *to appoint*). In the case of her male opponents the semantic profiling was quite different. Regarding the name Sarkozy, the most salient dimensions were the field of ACTION (*to sign*, *to hire*, *to act*, *to do*, *to depart*), and the field of OBJECTIVES (*to aim*, *to search*, *to covet*, *to want*, *to consider*, *to reckon*, *to wish*). The semantic profile of the name Bayrou also emphasized ACTION towards a goal, but in a more positive fashion than for the name Sarkozy (*to progress*, *to consolidate*, *to continue*, *to want*, *to offer*, *to make*); it further emphasized REBELLION (*to be unfavorable*, *to be against*, *to denounce*, *to disapprove*, etc.). A comparison of the three semantic profiles, we can we identify the gender stereotype defining the female versus male candidates reveals them defined according to the opposition ACTION (male) vs. COMMUNICATION (FEMALE). This opposition recalls the cliché of women being good in relating to people and men being better in taking steps and making decisions.

These results can be explained in a number of ways: by the discourse of the candidates, by their behavior (Royal and Sarkozy both played on sexual stereotypes), or by the journalists' perceptions of the politicians. It is impossible to identify the source of the difference-- but the difference is there.

From a semantic point of view, SEMY has revealed what we call the ‘deep structure’ of the names Bayrou, Royal, Sarkozy, as constructed in discourse. The TERMOSTAT software (Drouin 2003; Lesage *et al.* 1993), however, reveals another facet of the meaning, as it identifies the *specificities* of words found in the context of each name and confirm a difference for each name. This is the topic of the next section.

### 3.2 Antconc, KWIC abstracts and agency: power in politics?

With the concordance supplied by ANTCONC, we analyzed the KWIC abstracts (see section 1.3). To evaluate agency, we focused on the syntactic functions of the names, considering that Eckert and McConnell-Ginet argue that any NP in the position of subject in a sentence will be associated with power and control (2003, 52). We also evaluated the semantic functions to examine the actual *causality* of each NP, and restricted our study to agent and non-agent categories (Dowty 1991). This evaluation had to be done manually (Baider 2009), and we found that the syntactic choices made in journalistic discourse when referring to the presidential candidates revealed a partial dichotomy with the nouns (see Table 4).

|  | Sarkozy     |                     | Royal       |                    |
|--|-------------|---------------------|-------------|--------------------|
| <b>Subject</b>   | 158<br>(21) | 63.00 %<br>(13.30%) | 135<br>(48) | 46.55%<br>(35.60%) |
| <b>(Verb copula)<br/>(Passive voice)<br/>(Unaccusative Verb)</b> |             |                     |             |                    |
| <b>Object (direct)</b>   | 15          | 5.60%               | 14          | 4.83%              |
| <b>SN (<i>de</i> + Sarkozy or Royal)</b>                         | 37          | 14.75%              | 68          | 23.45%             |
| <b>SP</b>  | 41          | 16.33%              | 73          | 25.17%             |

Table 4. Syntactic roles for NPs Royal and Sarkozy drawn from KWIC abstracts (Baider & Jacquey 2010)

The noun Sarkozy functions more often as a subject than does the noun Royal. This contrastive study of syntactic choices used to describe a female versus a male candidate confirms a perception of the journalists conforming to stereotypes: the noun for the male candidate is, syntactically, in control of the verb, while the noun representing the female candidate is its subordinate. There was an even greater difference with regard to the semantic functions.

|   | Sarkozy<br>251 |                 | Royal<br>290 |                 |
|---|----------------|-----------------|--------------|-----------------|
| <b>Non-agent</b> = Theme, experiencer, location, cause, beneficiary | 114            | = 45.41%        | 212          | = 73,10 %       |
| <b>Agent</b>  | 133            | = <b>52,98%</b> | 74           | = <b>25,51%</b> |
| ?   | 4              | = 1,50%         | 4            | = 1,40%         |
|   |                | (99.93%)        |              | (99.98%)        |

Table 5. Thematic role of the noun *Sarkozy* (251 verbs) and the noun *Royal* (290 verbs) (Baider 2009)

The male candidate is constructed first as an agent, which concurs with SEMY findings. In contrast, the female candidate’s name most often played the role of an aim (*for* Royal), a place (*chez* Royal) or a relation (*de* Royal), which we categorized as cause. Moreover, the noun Royal was described twice as frequently as an experiencer, i.e., a person who feels, than as a person who controls the action described in the

sentence (Baider 2009). As a matter of fact, when both candidate names are in the same sentence, in 75% of cases the noun *Royal* is found in a clause embedded in the clause in which the noun *Sarkozy* is the subject, as the occurrences found in corpus show below (Baider 2009):

(1) Prenant de vitesse *Ségolène Royal*, qui doit débiter, lundi 25 septembre, un voyage au Sénégal, *Nicolas Sarkozy* a effectué, l'avant-veille, un déplacement de quelques heures à Dakar ;  
'Surprising Ségolène Royal, who will travel to Senegal on Monday 25 September, *Nicolas Sarkozy*, just a day ago made a short trip to Dakar'.

(2) *Nicolas Sarkozy*, qui battra *Mme Royal* 36 % contre 34 %, (...)  
'*Nicolas Sarkozy*, who would beat Ms Royal 36 % against 34 %, (...)

(3) *M. Sarkozy* croit que Mme Royal « le protège beaucoup ».  
'*M. Sarkozy* believes that *Ms Royal* "protects him a lot".'

The opposite is less likely to be found:

(4) Au moment où *Nicolas Sarkozy* s'apprête à rendre publique sa candidature (...) *Ségolène Royal* prend le large.  
'As soon as *Nicolas Sarkozy* is ready to publicly announce his candidacy (...), *Ségolène Royal* takes a trip'

We can also observe that the semantic prosody (Bednarek 2008) in all the occurrences above are consistently negative for *Royal*: *Royal* is surprised by *Sarkozy*'s quick move; whereas as soon as he announces his candidacy (an important move), she goes on a trip (a trivial endeavor)—here we also note the female candidate as the target of ironic observations (3) and (4).

The French press coverage of *Royal* during the 2007 presidential campaign may have been driven by and perpetuated a normative view of the sexes. However, the results may also be explained as much by the politician herself and her choices of words and personae. Therefore the linguistic profile could also be considered as an exception, but not the rule, given the results presented in the next section.

#### **4. Smaller candidates: specificity, power and gender in politics**

We conducted a second test to verify the validity of our previous hypothesis i.e. that studying newspaper political discourse might reveal a gender /power relationship established through linguistic choices. For this reason, we ran similar linguistic analyses on media coverage of less important elections with strong female candidates, investigating gender relevance when power is not as much at stake.

- In the 2007 presidential elections, we assessed the linguistic profiles of candidates with a smaller following, such as Voynet, Le Pen,
- In the 2011 socialist primaries we analysed the specificities of the data for Aubry and Hollande who belong to the same party.

##### *4.1 Qualitative reading, Semy and Antconc, the agency criteria*

An earlier study that focused on the minor candidates in the 2007 elections (Baider & Jacquy 2010) did not reveal the same asymmetry observed for *Royal* and *Sarkozy*. Using the same methodology we identified semantic profiles and agency for candidates of smaller parties, including (Ms) Voynet for the Greens, (Ms) Buffet for

the communist party and (Mr) Besancenot for the radical left-- all candidates who had no chance to win.

The lexical asymmetry found in relation to Royal did not hold true for the other female candidates who are part of our study. For example, with regard to male and female candidates in general, *negativity* was more a masculine trait than a feminine one, whereas in the study with Royal, *negativity* characterized the female candidate. Indeed, the male noun *Bové* was defined by the field of **Unlawfulness**; *Besancenot* by the field of **Discord** and *Le Pen* (Mr) by the field of **Concealment**. The three female politicians were characterised, as was Royal, by their relational personalities, with the field of **Advice** for *Buffet*, of **Moderation** for *Voynet* and of **Persuasion** for *Laguiller*.

As far as grammatical choices are concerned, the female Communist party leader Buffet and her newly appointed male opponent, Besancenot, the Revolutionary Communist party representative, served as a case in point. Both names were considered together, since the NP *Buffet* and *Besancenot* are found in the same syntagm 89% of the time when they are coordinated with another NP (such as in *Buffet et Besancenot*). At almost all levels of discourse our results contradict the figures obtained with the NPs Bayrou, Sarkozy and Royal. Indeed:

- The name *Besancenot* is used with another NP and therefore 'not autonomous' in 54% of the occurrences, whereas the name *Buffet* stands alone 35% of the times.
- The female candidate, Buffet, is twice as often an agent as her young male counterpart.
- On the lexical level, Besancenot, a Trotskyite, is also described as Alain Krivine's 'young protégé' (*jeune poulain*), Krivine being a well-known and longtime figure of the extreme left. The term 'protégé' reinforces his (just) burgeoning career. In contrast, the name Buffet co-occurs with the adjective *fort* (strong) and the noun *chef* (leader), signaling the long political career of the communist leader.

Even when both names occur in the same sentence, no striking pattern of hierarchy can be noted: both names are used in the same clause, both function as a subject or as part of a prepositional phrase.

The above-mentioned criteria are the only ones to differentiate the two discourses and they all favour the females candidates. In this we witness a situation where the stakes are not as high as those in a presidential election, and the same gender stereotypes in the political sphere are not apparent. For example, Buffet (the then communist party leader) was granted more agency than her young opponent Besancenot (the then extreme left leader). Is it perhaps that parity can be granted when there is not much to lose or to gain,<sup>12</sup> as Moores observed in 1997.? However, as mentioned before (Baider 2009; Baider & Jacquy 2010), the Sarkozy and Royal candidacies were well-known to be a somewhat exceptional case, since Royal and Sarkozy both played, and for the first time in French political history, on sexual stereotypes (Gingras 1996; Achin & Dorlin 2008; Coulomb-Gully 2012).

#### 4.2 Quantitative reading, Termostat, the specificity criteria

Although our data has been described already, we want here to reaffirm that we are investigating the socialist primaries in which one female and one male candidate were the final contenders. They were not the initial favorites, however; Dominique Strauss Khan (DSK) was clearly the favorite of the party until he was caught up in a scandal, which forced him to withdraw at the end of May.

---

<sup>12</sup> A tactic which is also followed by some parties when drawing up their list of candidates for elections: women are present in region where stakes are not high or where they are most likely to lose (Sineau 2010).

#### 4.2.1 Before the DSK affair

As Table 6 below illustrates, the most frequent terms associated with the two candidate names (as provided by TERMOSTAT) were terms found in articles focused on the political campaign (*candidate, first round, vote, campaign, party nomination, meeting*, etc.). They are therefore similar for all candidates.<sup>13</sup>

| Hollande    |           | Aubry       |           |
|-------------|-----------|-------------|-----------|
| Items       | Frequency | Items       | Frequency |
| candidat    | 290       | candidat    | 300       |
| primaire    | 248       | secrétaire  | 234       |
| président   | 171       | projet      | 212       |
| secrétaire  | 170       | primaire    | 204       |
| candidature | 153       | parti       | 161       |
| député      | 150       | candidature | 157       |
| campagne    | 121       | socialiste  | 109       |
| parti       | 108       | sondage     | 94        |
| sondage     | 97        | mai         | 90        |
| soutien     | 83        | juin        | 85        |
| projet      | 82        | maire       | 83        |
| maire       | 72        | député      | 82        |

Table 6: The most frequent items found for the NPs *Hollande* and *Aubry* in March-June 2011 in the six newspapers (Baider & Jacquey 2014)

However, if we look at the specificity criteria, and before the Strauss Khan fiasco, we see that the name Aubry is described extremely positively. Moreover, a close reading of newspaper discourse with TERMOSTAT revealed *no* blatant gender stereotyping in the six newspapers we used for our corpus. In fact, the opposite occurred: we observed reverse stereotypes. The discourse described the name Aubry as a dynamic and organised manager, and the name Hollande as an embattled politician busy dealing with conflict. In Table 7 below listing the specificities, we note three odd expressions typifying the candidates (i.e., after eliminating all the expressions common to both candidates).

In order to categorize the lexical units, first we used the interface of the morpho-syntactic and semantic criteria (denoting positive or negative values) :

- verbs (denoting action) are underlined,
- adjectives (describing positive or negative assessment) in italics,
- nouns (with a positive or negative connotation) in bold

We then grouped the lexical units according to their morphosyntactic categories.

Verbs, which can denote a more active and dynamic personality (or at least of a perception of a more dynamic person), are only found with the NP Aubry; none are recorded for Hollande:

Aubry: MA says (*dit Martine Aubry*), MA does (*fait Martine Aubry*), MA replies (*réplique Martine Aubry*).

Nouns referring to the candidates are more positive for Aubry than for Hollande who is referred to as a challenger, as a substitute, and in the expression ‘except Holland -- *tout sauf Hollande* (all except Hollande):

Hollande: deputy (*député*), president of the council, challenger, substitute, (*candidature de*

<sup>13</sup> TERMOSTAT would be useful for teaching French as a foreign language in order to spot the most current lexical fields when describing a certain event.

substitution), Dutch (*Hollandais*), except Hollande

Aubry: manager (*patronne*), rallying figure (*rassembleuse*)

Adjectives are found describing Hollande but not Aubry; however they are unfavorable (*anti-hollande*) or they describe the theme of his campaign, e.g., he will be a ‘normal candidate’; *normal* as adjective can be interpreted as a choice to describe Sarkozy, his political opponent, as not being a ‘normal’ president (because of his flashy life-style, as well as being far from ordinary citizens) or to describe Aubry as not being a usual candidate (because of her gender).

Hollande: normal (*président, présidence, candidature*), corrézien, anti-hollande,

| Hollande                            |        | Aubry                            |        |
|-------------------------------------|--------|----------------------------------|--------|
| N-grams                             | Spec   | N-grams                          | Spec   |
| <i>primaire socialiste</i>          | 155.97 | <i>primaire socialiste</i>       | 122.00 |
| <i>strauss-khanien(s)</i>           | 113.69 | <b>de Martine Aubry</b>          | 69.52  |
| <b>Candidature</b>                  | 67.11  | <i>projet socialiste</i>         | 64.95  |
| <i>président normal</i>             | 47.89  | <i>strauss-khanien(s)</i>        | 60.24  |
| <i>projet socialiste</i>            | 44.49  | <b>Patronne</b>                  | 49.52  |
| <i>ex-premier secrétaire</i>        | 52.38  | <b>emplois d’avenir</b>          | 45.02  |
| <i>candidat normal</i>              | 47.89  | <i>conseil politique</i>         | 35.93  |
| <i>projet socialiste</i>            | 44.49  | <i>après DSK</i>                 | 35.08  |
| <i>investiture socialiste</i>       | 44.03  | <u><i>que Martine Aubry</i></u>  | 35.08  |
| <i>présidence normale</i>           | 43.77  | <u><i>dit Martine Aubry</i></u>  | 35.08  |
| <b>Sympathisants</b>                | 41.72  | <b>Rassembleuse</b>              | 35.08  |
| <i>candidat déclaré</i>             | 39.32  | <i>Présidentiable</i>            | 34.89  |
| <i>humour corrézien</i>             | 38.73  | <i>égalité réelle</i>            | 34.06  |
| <b>Député</b>                       | 38.73  | <i>parti socialiste</i>          | 33.92  |
| <b>candidat de substitution</b>     | 34.11  | <i>bureau national</i>           | 32.79  |
| <b>Hollandais</b>                   | 32.54  | <i>candidat déclaré</i>          | 31.99  |
| <i>position de favori</i>           | 29.17  | <b>primaire de désignation</b>   | 28.93  |
| <b>Biceps</b>                       | 28.54  | <i>soubassement intellectuel</i> | 28.93  |
| <i>pro-dsk</i>                      | 28.13  | <i>convention nationale</i>      | 26.64  |
| <b>sauf Hollande</b>                | 28.13  | <b>emplois jeunes</b>            | 25.15  |
| <b>Présidentialité</b>              | 28.13  | <b>délai de décence</b>          | 25.03  |
| <i>front anti-hollande</i>          | 28.13  | <i>projet adopté</i>             | 24.74  |
| <i>expérience ministérielle</i>     | 27.81  | <b>texte de soutien</b>          | 23.31  |
| <b>Ralliement</b>                   | 25.82  | <b>coup de tonnerre</b>          | 22.75  |
| <i>contrat de génération</i>        | 24.33  | <i>jeunes socialistes</i>        | 22.40  |
| <i>Rocardiens</i>                   | 24.33  | <b>Partisans</b>                 | 21.68  |
| <i>stature présidentielle</i>       | 24.33  | <b>primaire d’octobre</b>        | 21.25  |
| <i>dépôt officiel</i>               | 24.33  | <u><i>Martine Aubry fait</i></u> | 21.25  |
| <i>opposant farouche</i>            | 24.33  | <u><i>réplique Aubry</i></u>     | 21.25  |
| <i>justice fiscal</i>               | 24.04  | <b>face à face</b>               | 21.25  |
| <i>referendum européen</i>          | 24.04  | <u><i>Martine Aubry a</i></u>    | 21.25  |
| <b>Soutien</b>                      | 23.29  | <i>thèmes économiques</i>        | 21.25  |
| <b>président de conseil général</b> | 22.68  | <b>coude à coude</b>             | 21.25  |
| <b>Challenger</b>                   | 22.65  | <i>image de rassembleuse</i>     | 21.25  |
| <i>adversaire socialiste</i>        | 21.79  | <i>candidature potentielle</i>   | 21.25  |

Table 7. The most specific n-grams associated with the NPs *Hollande* and *Aubry* in March-July 2011 in the six newspapers (Baider & Jacquey 2014)

We can also observe that the general tone of the specificities for Hollande are more negative than for Aubry (*adversaire, opposant farouche, sauf Hollande, front anti-Hollande*), a hostility which does not occur in Aubry specificities.

Our findings are corroborated by other studies. Marchand & Ratinaud (2012, 692) listed the most specific words as found in the socialist candidates *speeches*, finding that for Aubry out of 35 specific words, almost 40% are verbs, half of them denoting dynamism and actions such as *battre* ‘to beat’, *augmenter* ‘to rise’, *arrêter* ‘to stop’, *défendre* ‘to defend’, *fermer* ‘shut down’; in contrast only 12 % of the most specific



words identified for Hollande are verbs and they do not denote actions (*falloir* ‘must’, *avoir* ‘to have’, *évoquer* ‘to refer’). These same topoi characterized Sarkozy and Royal respectively in 2007.

#### 4.2.2 After the DSK affair

However, in summer 2011 a shift in the linguistic profiles could be seen, once it was clear that Aubry and Hollande were both favourites to represent the party against the incumbent president. Despite his lack of personality and the hostility he had earlier generated (as described in the section just prior), Hollande now seemed to have become the choice of the left. While only three months earlier, Aubry was the ‘*rassembleuse*’ (see Table 7 above), in the August/ September data, the picture had changed radically: *vote Hollande c’est pour éviter la fragmentation* ‘voting for Hollande will avoid fragmenting’; *il est mieux à même de rassembler la gauche* ‘he is better suited (than Aubry) to unite the left’. We must assume there had been a reversal in the party discourse, as we found it reflected in the journalist word choices: Hollande was thought more ‘apt’ to beat Sarkozy.

Why would Hollande suddenly be better suited to the job than Aubry? Maybe this choice could be explained in part by Hollande’s lead in the opinion polls? We could only conclude that, despite a positive journalistic profile, Aubry was never the favourite, neither in the polls, nor in the party. Despite his negative linguistic portrait, Hollande, in fact, actually led in the polls, and once Strauss Khan withdrew from the race, he became his party’s choice.

Our data presented here suggest that female and male politicians are likely to be treated equally by their party and / or in the French press—just so long as the stakes are not high. It seems that once they get higher, such as being potentially elected president, female politicians face negative discrimination. It is also possible to conclude that because journalists are linked to political parties, they write in a way to manipulate people, so that the party’s preferred candidate is chosen by the electorate. However as most studies of French political life have shown (Sineau 2010 *et passim*; Coulomb-Gully 2012; Bertini 2002; Gerber 2009; Adler 1993) the ‘chosen candidate’ in French politics,--whether chosen by the party or by the public-- is a gender issue.

### **Concluding remarks**

Through our study we have shown how extraction of discourses markers (mainly lexical and syntactic) using NLP technologies can address the issue of gender or political favoritism in journalistic discourse. We have also shown that conclusions drawn from qualitative studies may be supplemented, refined and nuanced by working the same data using specific computer software. We have shown that SEMY (based on definitions of the digital version of *Le Trésor de la langue française*) was able to offer a reading based on the most salient semantic features found in the contextual environment of the names Royal, Bayrou and Sarkozy. Finally, we clarified and supplemented the linguistic profile for each candidate found with SEMY using TERMOSTAT, as it works with specificities.

Even though these observational tools must always be improved, since the lists they provide require manual work, we highly recommend adding these tools when working on textual meaning, provided that the researcher has a good knowledge of the context in which the data have been produced.

To summarize:

- Manual studies are essential for syntactic-semantic research, since at present they are impossible automatically; such a qualitative reading also gives researchers a comprehensive picture of their data, a prerequisite for deciding the most appropriate interpretation. ANTCOINC provides the KWIC abstracts which greatly facilitate this evaluation.
- SEMY reveals recurring semantic features which cannot be detected with a superficial reading: it brings to light isotopes (Rastier 1987), which are the semantic characteristics recurrent in a text.
- TERMOSTAT reveals how these isotopes are realized and how different they are for each candidate.

These three methodologies, therefore, appear to be complementary and to work synergistically in the analysis of textual meaning. However, to draw broader conclusions from these data, and even insofar as gender is concerned, we must examine the other dimensions of the communicative event, namely, the journalistic text, as Schiffrin (2006) and Cotter (2011) have underscored. These other dimensions or factors are the ethnographically situated pragmatic arrangements that obey constraints, sociocultural values, and community-of-practice conventions (Cotter 2011: 2532). The most important in our case are the discourse practices (i.e., here the processes of text production in the French press) and the socio-cultural practices (i.e., here the French society). These must be taken into account in linguistic investigations of social power, since our focus is on the ‘connection between words and the world’ (Schiffrin, 2006, 333).

A discussion that considers both quantitative and qualitative analyses and the contexts of use (particularly social attitudes towards gender) will clarify the relations between ‘practitioner, text, and the shared world of all participants in the discourse, and what that can say about the role of women at any point in time in a particular sociocultural context’ (Cotter 2011, 2532). We therefore recommend the use of corpus linguistic tools as a valuable method in political and social science, and believe they could be a useful prerequisite for qualitative research in the digital humanities.

## References

Achin, Catherine & E. Dorlin (2008). Nicolas Sarkozy ou la masculinité mascarade du président. *Raisons politiques* 31: 19-46.

Adler, Laure (1993). *Les femmes politiques*. Paris: Seuil.

Alim, Samy & Reyes, Angela (2011). Introduction: Complicating race: Articulating race across multiple social dimensions. *Discourse Society* 22: 379-384.

Baider, Fabienne & Jacquey, Evelyne (2010). Substantive Dis-Embodiment, Syntactic Embedment. *Feminism, Femininity and Gendered Discourse* (eds. J. Holmes & M. Marra). Cambridge Scholar Press. 145-167.

Baider, Fabienne. (2009). Lexical change, discourse practices and the French press: Plus ça change, plus c’est la même chose? *Language Variation – European Perspectives II* (eds. S. Tsipalou, M. Karyolemu et P. Pavlou). Amsterdam: John Benjamins. 27-46.

Baider, Fabienne & Jacquey, Evelyne (2014). Data mining for Tracking Gender Bias: a Curse or a Blessing? The 2012 presidential duos. *Ecological and Data-Driven Perspectives in*

*French Language Studies* (eds H. Tyne *et al.*). Newcastle upon Tyne: Cambridge Scholar Press. 222-245.

Baider, Fabienne & Constantinou, Maria (2014). How to make people feel good when wishing ill: Golden Dawn and National Front Discourse, Emotions and Argumentation *2014 Yearbook of Corpus Linguistics and Pragmatics*, (ed. J. Tromero). Springer.

Barnes Julian & Larrivée Pierre (2011). Arlette Laguiller: Does the mainstay of the French political far-left enjoy linguistic parity with her male counterparts? *Journal of Pragmatics* 43:2501-2508.

Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory* 4/2: 119-139

Bergvall, Victoria. L.(1994). 'Cultural Projections in Constructed Linguistic Examples: Gender Representations in Introductory Texts, in M. Bucholtz et al. (eds.), *Cultural Performances*, Berkeley Women and Language Group, Berkeley, California, 15-27.

Bertini, Marie.-Josèphe (2002). *Femmes, le pouvoir impossible*. Paris:Pauvert/Fayard.

Brasart, Patrick (1998). Petites phrases et grands discours. *Mots. Les langages du politique* 40 : 106-112.

Caillat, Dominique (2015). Le recours au discours rapporté dans le débat présidentiel de l'entre-deux tours de 2007 : des usages sexués? *Les discours rapportés dans les discours médiatiques, une affaire de genre ?* (eds. F. Sullet-Nylander, M. Roitman, L. Rosier, S. Marnette et J.-M. Lopez). Stockholm: Stockholm University Press.

Clark, Kate (2005 [1998]). The linguistics of blame. *The Feminist critique of Language* (ed. D. Cameron). Routledge.

Cotter, C. (2011). Women's place at the Fourth Estate: Constraints on voice, text, and topic. *Journal of Pragmatics* 43: 2519–2533.

Coulomb-Gully, Marlène (2009). Présidentielles 2007 : Média genre et politique. *Mots* 90 (1) : 5-11.

Coulomb-Gully, Marlène (2012). *Présidente : le grand défi !* Paris: Éditions du Seuil.

Devriendt, Emilie (2015). Du discours rapporté dans deux séquences médiatico-dis, cursives de la pré-campagne présidentielle française 2012: candidats vs. candidates, une loi du genre? *Les discours rapportés dans les discours médiatiques, une affaire de genre ?* (eds F. Sullet-Nylander, M. Roitman, L. Rosier, S. Marnette et J.-M. Lopez). Stockholm: Stockholm University Press.

Dowty, D. (1991). Thematic proto-roles and argument selection, *Language* 67(2): 547-619

Drouin, Patrick (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology* 9 (1): 99-117.

Dupuy, P.-O. & Marchand ,P. (2013). Évolutions lexicales des débats de l'entre-deux-tours de l'élection présidentielle française. *Bulletin Vals-Asla (Bulletin suisse de linguistique appliquée)* 98 :105-128.

- Eckert, Penelope & McConnell-Ginet, Sally (2003). *Language and gender*. Cambridge: Cambridge University Press.
- Fairclough, Norman (1989). *Language and Power*. London: Longman
- Fairclough, Norman (1995). *Media Discourse*. London: Edward Arnold.
- Fracchiola Béatrice (2011). Politeness as a strategy of attack in a gendered political debate—The Royal–Sarkozy debate *Journal of Pragmatics* 43:2480-2488.
- Gerber, Gwendolyn L. (2009). Status and the gender stereotyped personality traits: Toward an integration. *Sex Roles* 61: 297–316.
- Gidengil, Elizabeth & Everitt, Joanna (2003). Talking Tough: Gender and Reported Speech in Campaign News Coverage. *Political Communication* 20: 209-232.
- Gingras A.-M, (1996). Les métaphores dans le langage politique. *Politique et Société* 30 : 159-171.
- Grzesitchak Mick, Jacquy, Evelyne & Valette, Mathieu (2007). *Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies*, ARCo'07 – Cognition, Complexité, [http://mathieu.valette.free.fr/Articles/Grzesitchak\\_Jacquy\\_Valette\\_2007.pdf](http://mathieu.valette.free.fr/Articles/Grzesitchak_Jacquy_Valette_2007.pdf)
- Goodwin, M.H. (2011). Engendering Children's Play: Person Reference in Children's Conflictual Interaction. *Conversation and Gender* (eds. S. A. Speer and E. Stokoe). Cambridge: Cambridge University Press. 250-271.
- Kahn, Kim Fridkin (1994). The distorted mirror: Press coverage of women candidates for statewide office. *Journal of Politics* 56 (1): 154-173.
- Kahn, Kim Fridkin (1996). *The political consequences of being a woman: How stereotypes influence the conduct and consequences of political campaigns*. Columbia University Press.
- Kanter, Rosabeth M. (1977). *Men and women of the corporation*. New York: Basic Books.
- Kitzinger, C. (2007). Is 'woman' always relevantly gendered? *Gender and Language* 1(1): 39-49.
- Kohl, Diane. (2011). The Presentation of "Self" and "Other" in Nazi Propaganda. *Psychology & Society* 7 – 26.
- Krieg, Planke A. (2011). Les "petites phrases" : un objet pour l'analyse des discours politiques et médiatiques. *Communication & Langages* 168 : 23 – 41.
- Kucera, Henry Francis, W. Nelson Carroll, John B. & Twaddell, W. F. (1967). *Computational Analysis of Present-Day American English*. Brown University Press
- Halliday, M.A.K. (2002). *The Collected Works of M.A.K. Halliday*. London: Continuum.
- Lazar, Michelle M. (ed.) (2005). *Feminist critical discourse analysis: Studies in gender, power and ideology*. Palgrave Macmillan.
- Le Bart, Christian (2010). Parler en politique. *Mots. Les langages du politique* 94 : 77-84.
- Lebart Ludovic & Salem André (1994). *Statistique textuelle*, Dunod.

Lesage, René, Price, Wilson, Bissonnette, Carole & Drouin, Patrick (1993). Enquête sur l'état d'utilisation des outils automatisés d'aide à la rédaction dans les organisations. *META* 38 (2): 367-389.

Macaulay, Monica & Brice, Coleen (1997). Don't touch my projectile: Gender bias and stereotyping in syntactic examples. *Language* 73 (4): 798-825.

Marchand, P. & Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 13-15 juin 2012, <http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Marchand,%20Pascal%20et%20al.%20-%20L%27analyse%20de%20similitude%20appliquee%20aux%20corpus%20textuels.pdf>

Marchand, P. (2009). Genre, style et attitude à l'égard du langage : tentative de diagnostic automatique sur un corpus politique. Journées d'études « Le style et ses modélisations » Université de Tours, 10-11 décembre 2009. [http://pascal-marchand.fr/IMG/pdf/PM\\_Tours.pdf](http://pascal-marchand.fr/IMG/pdf/PM_Tours.pdf)

Michard, Claire (1988). Some socio-enunciative characteristics of scientific texts concerning the sexes, the nature of the right. *A feminist analysis of order patterns* (ed. G. Seidel). Amsterdam: John Benjamins. 27-61.

Michard, Claire & Ribery, Claude (2008 [1987]). *Sexisme et sciences humaines Pratique linguistique du rapport de sexage*. Lille: Presses Universitaires du Septentrion.

Moore, Pamela M. (1997). Women in the presidential race: Arlette Laguiller and Dominique Voynet, 1995. *Women in French Studies* 5 : 281-292.

Mussolff, Andreas (2008). What can critical metaphor analysis add to the understanding of racist ideology? Recent studies of Hitler's anti-Semitic metaphors. *Critical Approaches to Discourse Analysis across Disciplines* 2 (2): 1-10.

Ochs, Elinor (1992). Indexing Gender. *Rethinking Context : Language as an Interactive Phenomenon*. (eds. A. Duranti and C. Goodwin). Cambridge: Cambridge University Press. 335-358.

Peck, Jennifer (2006). Women and promotion: The influence of communication style *Gender and Communication at Work* (eds. M. Barrett & M. Davidson). Hampshire: Ashgate. 50-66.

Penelope, Julia (1990). *Speaking Freely: Unlearning the Lies of the Fathers' Tongues*. New-York: Pergamon Press.

Rabatel, Alain (2015). La parole des politiques soumise à contre-enquête' représente-t-elle de la même façon la parole des hommes et des femmes politiques ? *Les discours rapportés dans les discours médiatiques, une affaire de genre ?* (eds F. Sullet-Nylander, M. Roitman, L. Rosier, S. Marnette et J.-M. Lopez). Stockholm: Stockholm University Press.

Rastier, François (1987). *Sémantique interprétative*. Paris: PUF.

Reutenauer, Coralie, Valette, Mathieu & Jacquy, Evelyne (2010). De l'annotation sémique globale à l'interprétation locale : environnement et image sémiques d' 'économie réelle' dans un corpus sur la crise financière. *Cognitica – Actes du colloque de l'Association pour la*

*Recherche Cognitive Arco '09 : Interprétation et problématiques du sens* (9–11 November 2009, Rouen), 29-39.

Reutenauer, Coralie (2009). Analyse et modélisation sémantique à partir de ressources lexico-sémantiques : résumé, *Texto !* (XIV),1, <http://www.revue-texto.net/index.php?id=2093>.

Roncarolo, Franca (2000). A woman at the quirinal: Thanks, but no thanks. The social construction of women's political agenda in the 1999 Italian presidential election. *European Journal of Women's Studies* 7 (1): 103-126.

Saïtta, Eugénie (2008). Les journalistes politiques et leurs sources. D'une rhétorique de l'expertise critique à une rhétorique du cynisme, *Mots. Les langages du politique* 87: 113-128.

Schiffirin, Deborah (2006). *In Other Words: Variation in Reference and Narrative*. Cambridge: Cambridge University Press.

Sinclair, John M. (1991). Words and phrases. *Corpus, Concordance, Collocation*. Oxford: OUP. 70–75.

Sineau, Mariette (2010). *La Force du nombre : femmes et démocratie présidentielle*. Paris : Ed. de l'Aube.

Shaw, Sylvia (2013). An Ethnographic Investigation into Gender and Language in the Northern Ireland Assembly (eds. I. Poggi *et al.*). *Political Speech 2010* LNAI 7688, pp. 39–53.

Sreberny-Mohammadi, Annabelle & Ross, K. (1996). Women MP's and the media representing the body politic. *Women in Politics* 49 (1): 103-117.

Sung, Matthew Chit Cheung (2013). Media Representations of Gender and Leadership: From a Discourse Perspective. *Brno Studies in English* 39 (1) DOI: 10.5817/BSE2013-1-5.

Thimm, Caja, Koch, Sabine & Schey, Sabine (2003). Communicating gendered professional identity. *The Handbook of Language and Gender* (eds. J. Holmes & Miriam Meyerhoff ). Oxford: Blackwell. 528–549.

Valette, Mathieu (2010). Propositions pour une lexicologie textuelle. *Zeitschrift für Französische Sprache und Literatur* 37 :171-188.

Valette, Mathieu & Rastier, François (2006). Prévenir le racisme et la xénophobie – propositions de linguistes. *Les langues modernes* 2 : 68-77.

van Dijk, Teung (2006). Discourse and manipulation. *Discourse & Society* 17(2): 359-383.

Widdowson, Henry. G. (2000). On the limitations of linguistics applied. *Applied Linguistics* 21 (1): 3-5.