Analysis of Fidel Castro Speeches Enhanced by Data Mining

Sergio Peignier*[1] and Patricia Zapata[2]

[1]Univ Lyon 1, INSA Lyon, INRA, BF2I, UMR0203, F-69621, Villeurbanne, France
[2]Carrera de Lingüística e Idiomas, Universidad Mayor de San Andrés, La Paz, Bolivia

Fidel Castro was a major Cuban politician of the twentieth century who influenced different left-wing regimes and political movements in Latin America and around the world. His ability to seduce the masses and captivate his audience relied to a large extent on his rhetorical abilities. Therefore, studying Castro's political speeches is a crucial step towards understanding his political success. Some previous studies have addressed this issue, mostly using small discursive samples and only a few speeches. However, using a small and possibly non-representative sample is likely to lead to biased results. To overcome this problem, this work was carried out on a large corpus of 1,018 speeches and 7,548,480 words, combining state-of-the-art data mining tools and the linguistic discourse analysis methodology proposed by Patrick Charaudeau. Combining both techniques, we provide here a more representative characterization of Castro's main discursive strategies.

Keywords: Discourse Analysis, Word-Embedding, Subspace Clustering, Hierarchical Clustering

1 Introduction

Fidel Castro came to power in 1959 as the leader of the Cuban Revolution, and governed Cuba until 2008 (e.g. Gott (2007)). During this period, he had a strong presence in the political arena, influencing different left-wing regimes and political movements in Latin America and around the world (e.g. Padmos et al. (2017)). His ability to seduce the masses relied, to an important degree, on his rhetorical skills. Indeed, according to Charaudeau (2009b), the ability of a politician to captivate people's attention depends to a large extent on his rhetorical capabilities. Thus, speech skills are crucial for a politician who wants to persuade an audience to join his cause, making such abilities

* Sergio.Peignier.Zapata@gmail.com

the finest political weapons. Therefore, studying Castro's political discourse is a crucial step towards understanding his political success.

According to Gee (2014), the discourse analysis field is divided into 'micro-communities', that develop their own methodologies based on particular paradigms. Even if different approaches tackle discourse analysis from different angles, they often share common terminologies and tools; that blur the frontiers between them and complexifies their categorization. Indeed, there is no single classification, and many taxonomies of discourse analysis methodologies have been created (e.g. Barry (2002), Maingueneau (2016)). Nevertheless, some general and important discourse analysis families have been identified by several authors. For instance, sociolinguistics (e.g. Hudson (1996)) studies the variants of language usage in a linguistic community; conversational analysis (e.g. Richards and Schmidt (1983)) focuses on talk-in-interactions; critical discourse analysis (e.g. Fairclough (2013)) studies how language establishes and reinforces power relationships within social groups; pragmatic analysis (e.g. Widdowson (1995)) studies the intentions of the audience and the speaker, as well as the context of the speech; lexicometric analysis (e.g. Pêcheux (1995), Wiedemann (2013)) is a computer-assisted family that can study a large corpus by characterizing its *structure* (e.g. vocabulary, grammar). To some extent, this technique has also been used to study discursive strategies by quantifying and interpreting the co-occurrences of some *pivot words* (e.g. words chosen ad hoc, most common words). Unlike the lexicometric family, most other methods generally extract, manually and systematically, all the underlying discursive strategies used by the politician in a corpus.

Since systematic studies are *time-consuming* and *complex* tasks, most previous systematic investigations have been conducted on small discursive samples only, containing few speeches. However, *small* and possibly *non-representative* samples may lead to biased interpretations, as shown extensively in statistics (e.g. Ellis (2010)). The family that is less impacted by this problem is the computer-assisted lexicometric one. So, at first glance, it might make sense to use *only* lexicometric methodologies. However, this sole paradigm cannot replace all the other ones, since each paradigm has been developed to address different scientific questions. Indeed, each family has its own interests and merits, and all are complementary tools for the discourse analyst.

In this work, we extend the recognized, non-lexicometric discourse analysis methodology created by Charaudeau (1995) by combining it with state-of-the-art data mining tools to study a corpus of 1,018 speeches and 7,548,480 word occurrences. For this purpose, our method uses word-embedding (e.g. Turney and Pantel (2010)) and subspace clustering (e.g. Kriegel et al. (2009)) to partition the vocabulary of the corpus into clusters of words sharing the same discursive context. Next, our method organizes intra-cluster words in dendrograms, applying hierarchical clustering. In this article, we refer to such structures as *dendrogram prototypical discourses*, or simply *DP-discourses*. The key intuition in the design of our method is to model the entire corpus, as a set of DP-discourses, and then study the most representative ones, using Charaudeau's discourse analysis methodology. In this context, our main research question could be stated as follows: Does the analysis of a large corpus, using a hybrid approach that combines a traditional discourse analysis methodology and data mining tools, provide new insights and a more representative characterization of Castro's discursive strategies?

In this paper, we used the latter approach to characterize the main discursive strategies used by Castro. Here we reveal that Castro presents himself as an authority, an expert, committed to his duties and identified with his audience. Moreover, he presents himself as a hero that protects people against several enemies, which are depicted as the source of all problems. In this context, he represents his audience either as a victim or as a hero fighting against the enemies. In addition, he frequently alludes to the economic and social progress of his country under his administration. All these elements aim at evoking strong feelings in his audience, such as heroism, pride, hope and fear. Finally, Castro uses elaborate and detailed descriptions to increase the veracity of his speeches.

The contribution of this paper is two-fold: On the one hand, we propose a data mining framework that models a large corpus as a set of representative DP-discourses, which can be studied easily by means of non-lexicometric methodologies. On the other hand, we provide a representative landscape of Fidel Castro's discursive strategies. Indeed, this work is the first to conduct a systematic study of more than 1,000 speeches and 7,500,000 word occurrences, using the non-lexicometric methodology proposed by Patrick Charaudeau.

The rest of the paper is organized as follows: Section 2 presents related works; Section 3 introduces Charaudeau's discourse analysis methodology; Section 4 describes our data mining framework; Section 5 describes the corpus collection and composition; Section 6 presents the discourse analysis of Castro's speeches themselves; Section 7 compares our main findings to the main conclusions of previous works; Section 8 concludes this paper by offering a summary, and perspectives for future research.

2 Related works

Given Castro's political importance, several previous works have studied his rhetorical abilities from different perspectives, using varied methodologies. The presentation of these works is divided into two parts: Section 2.1 presents earlier works based on non-lexicometric methods, while Section 2.2 focuses on lexicometric approaches.

2.1 Non-lexicometric approaches

The pioneering work of Joyner (1964) used a discourse analysis methodology based on Aristotelian principles of rhetoric to study three well-known speeches. Another pioneering work, Fagen (1965), identified frequent topics and discursive strategies and studied their interrelations in order to understand the mechanisms enhancing Castro's charisma. More recently, Nieto et al. (2002) have developed a *conversational analysis* method to study the emotions conveyed by Fidel Castro and Hugo Chávez during their first conversation that was broadcast on radio and television. Belisario (2010) combined techniques from *cognitive linguistics* and *critical discourse analysis* to study discursive strategies based on metaphors in three well-known Castro's speeches. Recently, Reyes (2011) combined *sociolinguistics* and *critical discourse analysis* to study the discursive roles assumed by a politician (e.g. narrator, interlocutor), using as sources all speeches delivered by Castro between the 15th of April and the 14th of August of 2005 (124,321 word counts in total). Considering corpus size, this work has been the most representative systematic study to date.

2.2 Lexicometric approaches

The lexicometric methodology has been used by Serge De Sousa to analyze the chronological evolution of Castro's speeches. In one of his first works, De Sousa (2009a) studied 42 speeches delivered by Castro each 26th of July, for the national day of Cuba, between 1959 and 2004. De Sousa (2009a) clustered these speeches in five historic periods, using two dimensionality reduction techniques: correspondence analysis, developed by Benzécri et al. (1973)) and the so-called 'analyse arborée' created by Luong and Mellet (2003). In order to study the temporal evolution of discursive topics, De Sousa (2009a) extracted the 42 most frequent words within each period, and compared the evolution of these lists of terms. De Sousa (2009a) presented chronological interpretations on his observations, taking into account historical events that characterized each period. Lately, De Sousa (2012) has extended this work, and obtained similar results, by considering all available speeches.

In a different work, De Sousa (2009b) used the entire corpus of Castro's speeches to study the evolution of the concept 'pueblo' [people] in the discourses. First, the author quantified the frequency of this term over time, showing that Castro used it more often between 1959 and 1964, during the early period of his rule. The author also tracked the evolution of the *semantic network* of 'pueblo', by extracting terms that had a significantly higher and lower co-occurrence with this term. De Sousa (2009b) showed that Castro conveyed an idealized representation of the people: conscious, confident, strong, proud and revolutionary.

Until now, no previous work has aimed at extracting systematically the different discursive strategies from a large corpus of Castro's speeches. While early systematic approaches relied on a small corpus, ranging from three to a few dozens of speeches, lexicometric approaches have considered many speeches but only studied the evolution of the frequency and the semantic network of a few terms.

3 Discourse analysis methodology

The discourse analysis presented in this article is based on the so-called semio-pragmatic methodology, a non-lexicometric approach developed by Charaudeau (1995). The semio-pragmatic approach is a well-recognized methodology that has introduced some important founding principles into the field (Weizman, 2008), and Charaudeau is recognized as one of the most representative authors of the French School of discourse analysis according to Weizman (2008). This methodology is based on the Aristotelian classification of the art of rhetoric, which consists of three families of discursive strategies, called Ethos, Pathos and Logos. In this section, we present the major strategies of these three families, describing their underlying objectives, i.e. their expected impact on the audience. Nevertheless it should be noted that the outcome of a given discursive strategy, i.e. the audience reaction, may differ from the expected outcome, which also depends on external circumstances. Indeed, a discursive strategy can be said to have only a potential influence on a specific audience under particular circumstances. The discourse analysis methodology presented hereafter does not aim at characterizing the reaction of the audience; it only aims at identifying underlying discursive strategies.

3.1 Ethos

Ethos strategies allow the speaker to build his discursive identity, they strengthen his credibility and they enable the identification between the speaker and his audience. Hereafter are described the main mechanisms from the Ethos family that were identified by Charaudeau (2009a).

Discursive identity: A politician alternates between three discursive identities: The *"Me" identity*, when he speaks only on his behalf, using the first person singular; the

"Me-Us" identity, when he also speaks on the audience behalf, using the first person plural; and the *"Me-Spokesman" identity*, when he speaks on behalf of a doctrine or ideal.

Embody a credible character: A politician reinforces his legitimacy, by personifying credible characters, such as: 1) a *leader*, who imposes his decisions and emphasizes his institutional *authority of power*; 2) an *expert*, who exhibits his analytical competences and knowledge of a given topic; 3) a neutral *witness*, who removes from his speech any indications of personal judgment; 4) a person *committed* to his ideals, who vehemently defends his ideas as unquestionable truths; 5) a *charismatic friend or relative*, who has a strong identificatory relationship with his audience.

3.2 Pathos

Pathos strategies aim at persuading the audience by bringing out feelings and passions. Within this category, Charaudeau (2008, 2011) identified three major mechanisms, which are presented hereafter.

Recruitment process: This strategy aims at leading the audience to accept the speaker's project willingly. To do so, the speaker refers to classic positive values, such as social welfare (e.g. freedom, justice, security); national or regional belonging (e.g. nationalism, regionalism); religious, ethnic or ideological belonging; development, economic growth and technological progress; and moral values (e.g. honesty, commitment).

Rhetoric of effects: The goal of this mechanism is to stir feelings and passions in the audience, in order to predispose people to share the speaker's point of view. Indeed, it has been extensively shown in the literature that emotions can have an important influence on decision-making (e.g. Janis and Mann (1977), Schwarz (2000)). In the context of discourse analysis, the study of the rhetoric of effects mainly focuses on basic emotions that were reported in the field of cognitive sciences (e.g. Ekman (1992), Lövheim (2012), Plutchik (2001)): Usually the speaker aims at provoking feelings such as threat, fear, compassion, hope and pride. To do so, politicians may use the *dramatization* strategy, which consists in telling dramatic life stories that involve several characters that the audience can identify with or reject.

Triadic scenario: Politicians' speeches are often organized as a *triadic scenario*, which contains the following three elements: 1) A current or latent *disastrous social situation* is described by the speaker to induce the audience to a state of anguish and lead the public to speculate about the origin of the problems. 2) An *enemy* is pointed out by the politician as the cause of all problems. Enemies are either clearly identifiable (e.g. political party) or vague entities (e.g. ethnic groups), and they are commonly embodied by the speaker's political adversaries. 3) A *hero* is proclaimed by the speaker as society's savior and protector against enemies. Heroes are either abstract entities (e.g. social classes) or real persons (e.g. the speaker himself). This mechanism proposes a seductive imaginary where the audience is both the hero and the main beneficiary.

3.3 Logos

The Logos strategies are used to convince the audience through logical reasoning and argumentation. According to Charaudeau (2005), Logos is used less often than Ethos and Pathos in the context of political speeches. Indeed, for politicians it is less important to explain concepts in a logical way than obtaining the audience's support by using the most efficient strategy. Moreover, in the political context, the goal of

Logos is limited to increasing the truthfulness of the speech. The main Logos strategies identified by Charaudeau (2005) are described hereafter.

Singularization and essentialization: The *singularization* strategy aims at reducing the number of ideas exposed in the speech, keeping the attention of the audience focused on a few concepts. The *essentialization* strategy represents complex concepts by using only a few words. Once a concept has been essentialized, the audience does no longer need to reflect on it to make sense of it, which reduces the audience's mental effort. Both strategies are often combined to form a so-called *formula* strategy. According to Charaudeau (2005), a formula produces a strong feeling of evidence and attraction in the audience. Similar strategies called *slogans* concentrate entire ideas in catchy sentences, reminiscent of proverbs that seem to convey an absolute truth.

Self-evidence and causal arguments: Politicians often use *simple causal reasoning* to persuade the audience. They often build causal arguments upon values and beliefs that are deeply rooted in the minds of the majority of the audience. In order to enhance the causal arguments, politicians often present such values as being *self-evident assumptions*, i.e. known beforehand and accepted by everyone.

Analogies and detailed descriptions: To increase the veracity of their speeches, politicians often use *analogies with the past*, making reference to important historical characters or events. Finally, politicians include detailed descriptions and narrations to enhance the veracity of their speeches.

4 Dendrogram prototypical discourses

4.1 Founding principles

According to Harris (1954) and Rubenstein and Goodenough (1965), words in natural languages are structured within linguistic environments (e.g. sentences, paragraphs), and in this context, words having similar meanings tend to share similar contexts. This assumption, known as the *distributional hypothesis*, suggests that a corpus is often constituted by several discursive contexts, each one being a set of extended linguistic environments, conveying similar/related concepts and topics. Although this theory emerged in linguistics as early as 1954, it has recently received an increasing attention in many other fields such as in cognitive sciences (e.g. McDonald and Ramscar (2001)) and natural language processing (e.g. Mikolov et al. (2013a)). This hypothesis is the founding principle of our approach.

Our method aims at modeling a large corpus as a set of so-called DP-discourses and then studying them as prototypical speeches. To do so, the key step consists in building clusters of words sharing similar discursive contexts. This was achieved using word-embedding and subspace clustering, but other data-mining techniques could be used. Then, intra-cluster words were represented as *dendrogram prototypical discourses (DP-discourses)*, using a hierarchical clustering algorithm. Finally, DP-discourses have been studied using Charaudeau's methodology; they could possibly be analysed using other discourse analysis approaches.

4.2 Vector Space Modeling

This step aims at representing the corpus' different words as real-valued numeric vectors, building a Vector Space Model. For this purpose we used Word2Vec, a well-known word embedding algorithm based on neural networks, developed by Mikolov et al. (2013a). Word2Vec Vector Space Models are able to capture the contextual and the semantic relationship between words, such that words appearing in the same context tend to have similar vector representations. Word2Vec has two major architecture: skip-gram and CBOW. According to Mikolov et al. (2013a), skip-gram outputs better representations for small datasets, while both architectures provide similar results for large datasets; regarding runtimes, CBOW tends to be faster than skip-gram. Since in this work we are dealing with a large dataset, we decided to use the CBOW architecture, with a negative sampling technique, as detailed hereafter.

Formal definition Let a textual corpus be a list $(w^{(1)}, w^{(2)}, \dots, w(n))$, and let V_W denote its vocabulary, such that $\forall i \in \{1, \dots, n\}, \quad w^{(i)} \in V_W$. The frequency of a word $w \in V_W$ is simply the number of times it appears in the corpus. The context $c^{(i)}$ of a word $w^{(i)}$ is defined as the list of neighboring words, within in a window of size l: $c^{(i)} = (w^{(i-l)}, \dots, w^{(i-1)}, w^{(i+1)}, \dots, w^{(i+l)})$. The set of contexts in the corpus is called V_C. Each word $w \in V_W$, and each context $c \in V_C$ are represented respectively by vectors $x \in \mathbb{R}^D$ and $z \in \mathbb{R}^D$, where D is the dimensionality of the Vector Space Model. A word-context pair $\langle w, c \rangle$ exists in the corpus, if c is the context of w, at least once. The probability that the pair $\langle w, c \rangle$ exists in the corpus is denoted $P(exists|w, c)$, and the probability of the complementary event is simply $P(\overline{exists}|w, c) = 1 - P(exists|w, c)$. In Word2Vec, this probability distribution is approximated using the corresponding vector representations x and z, as follows: $P(exists|w, c) \approx p(x, z) = 1/(1 + e^{-x^\top z})$. Word2Vec relies on a negative sampling technique: its neural network is trained to learn the vector representations that allow to discriminate a world w from a set of k randomly drawn words denoted $\tilde{w}$, only using its context c. This is achieved by maximizing $log(P(exists|w, c)) + k \times \mathbb{E}(log(P(\overline{exists}|\tilde{w}, c)))$. Indeed, this expression aims at maximizing $P(exists|w, c)$ for existing pairs, while minimizing $P(exists|\tilde{w}, c)$ for unexisting ones (i.e. maximizing $P(\overline{exists}|\tilde{w}, c)$). In terms of word and context representations, the corresponding objective function formalization is: $\mathcal{L}(x, z) = log(1/(1 + e^{-x^\top z})) + k \times \mathbb{E}(log(1 - 1/(1 + e^{\tilde{x}^\top z})))$. To maximize $\mathcal{L}(x, z)$, the algorithm updates the word and context representations x and z for each example, using stochastic gradient descent.

Parameter setting The Word2Vec implementation available in the Gensim Python library (Rehurek and Sojka, 2011) was used with the following parameter setting: The Vector Space Model dimensionality was set to $D = 300$, the context window size to $l = 5$, the number of negative samples to $k = 5$, and the number of iterations across the entire corpus to $NbIter = 5$.

4.3 Subspace clustering

Once the Vector Space Model was built, we clustered its word vector representations. The aim of this step was to find groups of words sharing the same discursive context, to analyze them separately. This task could have been achieved using any traditional clustering technique (Jain et al., 1999); however Aggarwal et al. (2001) have shown that traditional data mining algorithms struggle in high dimensional spaces, such as the 300 dimensional Vector Space Model generated by Word2Vec. To overcome this problem, an alternative is to use subspace clustering. This data mining task is recognized as being more general than clustering, since it does not only search groups of similar objects but also detects the subspaces where similarities appear. In this work, we used the recent subspace clustering algorithm called SubCMedians, designed by

Peignier et al. (2018). This technique, based on a K-medians paradigm, groups data instances around centers, and updates the coordinates and the subspaces of the centers, to minimize the distance to their closest data objects, using stochastic hill climbing. The clustering procedure can be stated more formally as follows:

Formal definition Let X denote the set of vector representations, such that $x \in \mathbb{R}^D$ represents word $w \in V_W$, and D denotes the Vector Space Model dimensionality. Let $\mathcal{M}$ denote the set of centers built by SubCMedians, such that $m_i \in \mathcal{M}$ is defined in its own subspace $\mathcal{D}_i$. Let $dist(x, m_i)$ be the distance between x and m_i; in SubCMedians, $dist(x, m_i)$ corresponds to the Segmental Manhattan distance (Aggarwal et al., 1999), an extension of the Manhattan distance, that allows to deal with vectors defined in different subspaces. Each vector $x \in X$ is assigned to its closest center $m_i \in \mathcal{M}$, and the corresponding distance between them is termed the Absolute Error $AE(s, \mathcal{M}) = min_{m_i \in \mathcal{M}} dist(x, m_i)$. The objective of SubCMedians is to find the centers $\mathcal{M}$ that minimize the Sum of Absolute Errors $SAE(X, \mathcal{M}) = \sum_{x \in X} AE(x, \mathcal{M})$. Once a suitable set of centers $\mathcal{M}$ has been produced, each vector $x \in X$ is assigned to its closest center, together with its corresponding word w; such that $\mathcal{C}_i = \{\langle w, x \rangle, \ldots\}$ denotes the cluster of word-representation pairs associated to center $m_i \in \mathcal{M}$. This assignment step defines directly a clustering model of words and vector representations.

Parameter setting Peignier et al. (2018) provided a simple and effective default parameter setting procedure for SubCMedians: the user simply provides a suggested number of clusters $NbExpClust$, and the algorithm automatically adapts the number of clusters and the sizes of subspaces. In this work, we set this parameter to $NbExpClust = 2$, which turned out to be sufficient to build a satisfactory subspace clustering model, as shown in Section 6.1.

4.4 Dendrogram-based intra-cluster words representation

The pairs of words-representations belonging to each subspace cluster were organized in dendrograms, using the traditional bottom-up hierarchical clustering algorithm developed by Sokal (1958). The aim of this step was to provide an interpretable structure of the intra-cluster words, to study them using Charaudeau's methodology. Other visualization techniques could also be used for this purpose[1].

Formal definition Let $\mathcal{C} = \{\langle w, x \rangle, \ldots\}$ be a cluster of word-representation pairs. Initially, each pair $\langle w, x \rangle \in \mathcal{C}$ is considered as an isolated group $\{\langle w, x \rangle\}$. Then, at each iteration, the algorithm merges the two closest groups, considering distances in the Vector Space Model. Iteratively the number of groups decreases, until every pair belongs to the same cluster. Then the algorithm stops, and the hierarchical arrangement of groups produced by the algorithm is represented as a dendrogram. Since the construction of the hierarchical structures is based on the distances between vector representations, the corresponding dendrograms represent the contextual and semantic relationship between intra-cluster words.

Parameter setting The hierarchical clustering algorithm has two major meta-parameters. The first one is the distance used to compare individual vector rep-

[1] For instance, t-SNE plots (Maaten and Hinton, 2008) seem an interesting alternative to dendrograms, and should be tested in future works.

resentations. In this work, two vector representations ${}^u x$ and ${}^v x$, were compared using their Manhattan distance $||{}^u x, {}^v x||_1$. The second meta-parameter corresponds to the so-called linkage method, which is used to assess the similarity between two groups u and v, before merging them. Here we used the Complete linkage method, which considers the maximum distance between elements of each group, as a similarity measure: $distance(u, v) = max(\{||{}^u x, {}^v x||_1 : \langle {}^u w, {}^u x \rangle \in u, \langle {}^v w, {}^v x \rangle \in v\})$.

In this work, we tested nine meta-parameter configurations, combining three classic similarity measures: the Manhattan distance, the Cosine similarity, and the Euclidean distance, and three well-known linkage methods: Average, Complete, and Single linkage. In practice, for each one of the nine possible configurations, we used the hierarchical clustering package from SciPy Python library (Jones et al., 2019) to build the dendrograms and partition their main branches. While analyzing these structures, a common practice consists in studying first each branch independently and then combining the different interpretations to get an overall understanding. In this context, it is preferable to deal with dendrograms such that: 1) the words within each branch are densely packed together, forming groups with low intra-cluster dispersion; 2) the branches are well separated from each other, forming a partition with a high inter-cluster dispersion. This characteristic of a clustering structure is captured by a well-known clustering quality measure: the Variance Ratio Criterion (Caliński and Harabasz, 1974). This measure is simply the average ratio between the inter-cluster and the intra-cluster dispersion, and higher scores are obtained by more interpretable dendrograms. In order to choose the most suitable configuration, we have computed the average Variance Ratio Criterion of the dendrograms, obtained using each one of the nine configurations. According to the results, depicted in Table 1, among these nine configurations, the combination of Manhattan distance and Complete Linkage method, obtained the higher Variance Ratio Criterion, and hence this setting was chosen as the one providing the most interpretable dendrograms.

	Manhattan	Euclidean	Cosine
Complete Linkage	**10.52**	9.70	9.10
Average Linkage	5.09	5.02	6.98
Single Linkage	4.23	4.13	3.37

Table 1: Average Variance Ratio Criterion for nine combinations of similarity measures (columns) and aggregation techniques (rows).

5 Dataset

Before focusing on the discourse analysis, it is important to describe the corpus itself. In this section we discuss the textual origin of the speeches, the data collection procedure, as well as basic cleaning and pre-processing steps that were applied to the corpus.

5.1 Textual origin

The characterization of the speeches' textual origin and composition mechanism raised two important questions: 1) Did Castro write and/or prepare his speeches himself or did a specialized team prepared them for him? 2) Were the speeches improvised or were they prepared beforehand?

To the best of our knowledge, only one interview carried out by Ramonet (2010), has addressed these questions. According to this interview, Castro himself affirmed that some of his speeches were carefully written, while others were more or less improvised on the spot. In this context, Castro also highlighted the differences between the two kinds of speeches and he affirmed that written speeches may diminish the ability of the speaker to modulate his tone and tend to be less emphatic than improvised ones. Consequently, a reasonable expectation is that our dataset contains both kinds of speeches.

In the same interview, Castro declared that he had never been satisfied by speeches prepared by his collaborators and always ended up preparing his own speeches. Nevertheless, even if we imagine the speeches to have been prepared by a dedicated team of collaborators, it is reasonable to expect that the speaker would have read and validated the speeches and their underlying discursive strategies.

This work does not aim at classifying and analyzing the speeches according to their putative textual origin, which would be an interesting research subject on its own. Instead we have approached the analysis of Castro's speeches in a broad sense, regardless of their degree of preparation and textual origin.

5.2 Data collection

The corpus that has been analyzed in this work has been downloaded from a dedicated official website of the Cuban government.[2] This website hosts a large collection of documents produced by Fidel Castro, including speeches, interviews, essays and letters. Moreover, this website also hosts the translations of these documents into different languages. In this work, we decided to focus specifically on speeches in Spanish, filtering out interviews, letters, essays, and translated texts. This step aimed at avoiding possible biases by preventing the inclusion of strategies belonging to other kinds of enunciation, delivered through other channels of communication. In practice, data collection was ensured by Python web-scraping custom scripts, relying on the urllib2 (Van Rossum and Drake, 1995) and the BeautifulSoup (Richardson, 2019) Python libraries, while data filtering and cleaning were facilitated by Python custom scripts using the re (Van Rossum and Drake, 1995) and nltk (Bird et al., 2009) Python libraries. Finally, the corpus underwent a comprehensive manual verification.

5.3 Pre-processing

Once the corpus had been downloaded and cleaned, we applied stop-words filtering, a pre-processing step that excluded from the study extremely common words conveying too little semantic information. Even if stop-words filtering is a very common pre-processing step in natural language processing, it can be delicate to use this pre-processing in the context of word embedding. On the one hand, Mikolov et al. (2013b) have shown that a related procedure that aims at massively subsampling very frequent words can reduce the runtimes and improve the word vector representations of less frequent words. On the other hand, Agarwal and Yu (2009) have shown that removing stop words that are actually carrying semantic information (i.e. words linked to some specific contexts or negations) may lead to a significant quality drop of the word representations. In order to reduce runtimes and improve the representations of less frequent words while avoiding issues related to the exclusion of actually meaningful stop words, we carefully chose a curated stop words list with only 34 demonstrative adjectives, indefinite and definite articles. Stop words filtering was facilitated by Python custom scripts using the nltk (Bird et al., 2009) library.

[2] http://www.cuba.cu/gobierno/discursos/

In total, the corpus contains 1,018 discourses in Spanish, 7,548,480 word occurrences, 4,161,729 not-stop words occurrences and a vocabulary of 6,453 distinct not-stop words.

6 Results

For the sake of reproducibility and completeness, the Word2Vec vector representations, the clustering memberships, and all the DP-discourses are available on a dedicated web-page[3] and the software is available on a GitLab repository.[4]

6.1 Quantitative cluster assessment

Clustering structure Using the aforementioned methodology, the SubCMedians algorithm automatically extracted 26 clusters, thus adapting its number of clusters to the dataset without being restrained by the weak number of expected clusters parameter setting (here $NbExpeClust = 2$). For each cluster, we computed its absolute and relative vocabulary size, as well as the absolute and relative number of word occurrences. As shown in Figure 1, the 26 clusters have different sizes, and the four largest clusters gather close to 60% of the number of words and almost 50% of the corpus vocabulary. Hence, considering only these four clusters seemed sufficient to perform a representative discourse analysis study.

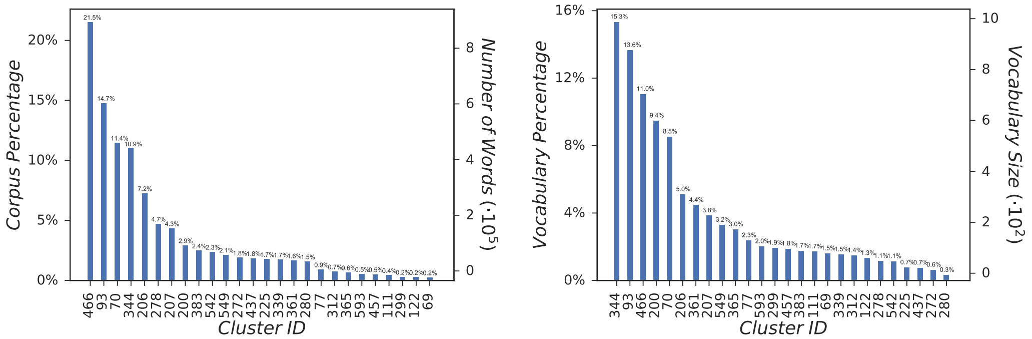

Figure 1: Absolute and relative number of word occurrences (right) and vocabulary size (left) per cluster.

Motivation for clustering assessment This work relies on the assumption that clusters represent discursive contexts, which allows their corresponding DP-discourses to be studied as prototypical speeches. According to the definition of discursive contexts presented in Section 4.1, consecutive words are likely to belong to the same discursive context. So, if clusters capture discursive contexts, consecutive words should also tend to belong to the same cluster, i.e. observing the same cluster membership for consecutive words should be mutually dependent events. Before studying the underlying discursive strategies embedded in the corresponding DP-discourses, we assessed whether our clusters exhibit this statistical property. To do so, we relied on the Weighted Point-wise Mutual Information measure, between the cluster memberships of consecutive words across all speeches, as detailed hereafter.

[3] https://analisisdiscursosocialistalatinoamerica.github.io/castro.

[4] https://gitlab.com/speignier/dendrogram_prototypical_discourses

Weighted Point-wise Mutual Information The Point-wise Mutual Information (PMI), as defined by Church and Hanks (1990), is a measure that quantifies the mutual dependence between two outcomes y and z, of two random variables Y and Z. The PMI is the logarithm of the ratio between the joint probability $p(Y = y, Z = z)$ and the distribution assuming independence $p(Y = y) \times p(Z = z)$. The variant called Weighted Point-wise Mutual Information (WPMI) simply weights the PMI by the joint probability: $WPMI(y, z) = p(Y = y, Z = z) \times \log\left(\frac{p(Y=y,Z=z)}{p(Y=y)p(Z=z)}\right)$. The WPMI between independent events is equal to zero this measure is positive for mutually dependent events, and it is negative for events that mutually exclude each other.

Intra and inter-cluster WPMI Let $C^{(t)}$ and $C^{(t+1)}$ be two discrete random variables, which model the cluster membership of two consecutive words from the corpus. Each random variable has K possible outcomes, denoted $\{c_1, c_2, \dots, c_K\}$, and each outcome corresponds to one of the existing clusters ids (here $K = 26$). The $WPMI$ between the cluster memberships c_i and c_j of consecutive words is defined as follows:

$$WPMI(c_i, c_j) = p(C^{(t)} = c_i, C^{(t+1)} = c_j) \times \log\left(\frac{p(C^{(t)} = c_i, C^{(t+1)} = c_j)}{p(C^{(t)} = c_i) \times p(C^{(t+1)} = c_j)}\right)$$

Moreover, let $IntraClustWPMI$ and $InterClustWPMI$ denote respectively the sets of intra-cluster and inter-cluster WPMI values. More precisely, $IntraClustWPMI$ and $InterClustWPMI$ are simply the sets of WPMI values of consecutive words having the same ($IntraClustWPMI = \{WPMI(c_i, c_i), \dots\}$) or different cluster memberships ($InterClustWPMI = \{WPMI(c_i, c_{j \neq i}), \dots\}$). Let the sums of elements in $IntraClustPMI$ and $InterClustPMI$ be denoted as $IntraClustMI$ and $InterClustMI$ respectively. These values correspond to the intra-cluster and the inter-cluster Mutual Information measures.

Probabilities estimation Let $\#(c_i, c_j)$ be the frequency of consecutive non-empty words belonging to cluster c_i and c_j. In practice, the joint probabilities $p(C^{(t)} = c_i, C^{(t+1)} = c_j)$ can be estimated by $\frac{\#(c_i,c_j)}{\sum_i \sum_j \#(c_i,c_j)}$, and the marginal probabilities $p(C^{(t)} = c_i)$ and $p(C^{(t+1)} = c_i)$ can be estimated by $\frac{\sum_j \#(c_i,c_j)}{\sum_i \sum_j \#(c_i,c_j)}$ and $\frac{\sum_i \#(c_i,c_j)}{\sum_i \sum_j \#(c_i,c_j)}$ respectively.

Intra and inter-cluster WPMI comparison The $WPMI$ values were estimated using the previous method, and then the results were organized in a matrix. The rows and the columns of the matrix represent the possible outcomes of the random variables $C^{(t)}$ and $C^{(t+1)}$, respectively, and they are labeled accordingly, so $WPMI(c_i, c_j)$ is located in the row with label c_i, and column with label c_j. As depicted in Figure 2, the highest $WPMI$ values from this matrix correspond to the intra-cluster $WPMI$ for clusters 70, 93, 344 and 466, suggesting that intra-cluster $WPMI$ measures are higher than inter-cluster ones. Moreover, the intra-cluster Mutual Information $IntraClustMI = 0.102$ is revealed to be higher than the inter-cluster $InterClustMI = -0.031$. In order to determine whether $IntraClustWPMI$ and $InterClustWPMI$ follow the same distributions, we applied the non-parametric Mann–Whitney U test. This test resulted in a Mann–Whitney U statistic equal to 14,211 and a very low p-value equal to 1.8×10^{-09}. Consequently, intra-cluster WPMI are significantly higher than inter-cluster ones, which supports the hypothesis that clusters represent discursive contexts, allowing us to proceed to the discourse analysis of the corresponding DP-discourses.

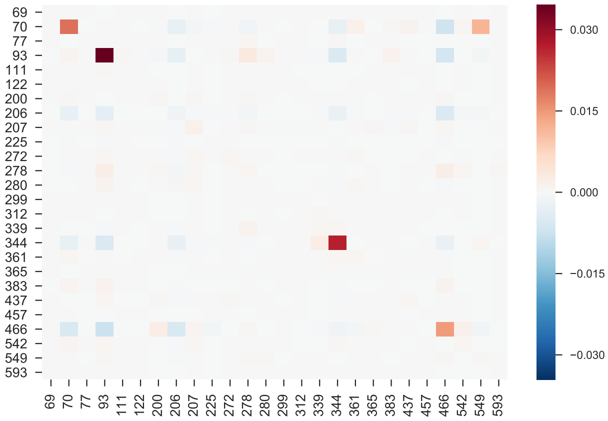

Figure 2: Matrix representing the *WPMI* measures between the cluster memberships of consecutive words. The value $WPMI(c_i, c_j)$ is located along row labeled c_i and along column with label c_j. A strong mutual dependence between events is depicted in red, while mutually exclusive events are depicted in blue.

6.2 Discourse analysis

Regarding the previous results, we focused on the four biggest clusters, with highest *WPMI* values, i.e., clusters 466, 344, 93 and 70. As explained in Section 4.4, the corresponding intra-cluster words were organized in DP-discourses, and analyzed using Charaudeau's methodology.

Castro's marxist revolutionary project (DP-discourse 466) This DP-discourse has a vocabulary of 713 words and 895,982 occurrences. The percentage of vocabulary size and the number of occurrences expressed in this DP-discourse are equal to 11.05% and 21.53% respectively. Analyzing this DP-discourse, we can infer that Fidel Castro's speeches often depict a revolutionary Marxist project for the future. To captivate his public, Castro characterizes his plan as being strongly tied to moral values, nationalism, social welfare and people. In this context, Castro summons his audience to join the project and also to defend it. To do so, he mainly uses recruitment processes and so-called 'rhetoric of effect' strategies. A DP-discourse representing the 100 most frequent terms from cluster 466 is illustrated in Figure 3.

Plan for the future: This DP-discourse is characterized by the presence of common nouns, and verbs in future and conditional, that evoke a forthcoming project and its near fulfillment (e.g. 'futuro' [future], 'adelante' [forward], 'será' [will be]). Different terms show that Castro depicts his plan as a difficult task (e.g. 'dificultades' [difficulties]), but nonetheless achievable (e.g. 'esfuerzos' [efforts]). Moreover, there are verbs in subjunctive and conditional, suggesting that this project is presented as a major aspiration of Castro and his public. Different verbs of obligation (e.g. 'seamos' [we must be]), in the first and third person, suggest that Castro presents the plan as an obligation. Therefore, Castro instructs Cuban people to join his project, and in doing so, he exhibits an Ethos of authority and commitment to this task. Moreover, since the notion of a future plan tends to convey an underlying idea of progress, Castro captivates his audience using a recruitment process strategy. Finally, by evoking the construction of a better future, Castro seeks to provoke in his audience strong feelings, such as hope and pride. This approach corresponds to the rhetoric of effects.

Marxist revolution: Castro's project is actually the consolidation of a Marxist popular revolution. Indeed, this DP-discourse is characterized by a large and rich lexicon reflecting a revolutionary Marxist imaginary (e.g. 'revolucionarios' [revolutionaries], 'socialismo' [socialism]), which shows the importance that Castro assigns to this topic. This strategy corresponds to a process of recruitment based on ideological belonging. Close to this lexical group, there are several nouns, evoking abstract social welfare topics, nationalism and mainly moral values (e.g. 'justicia' [justice], 'moral' [moral], 'patria' [homeland]). Hence, Castro includes, in the description of his plan, moral precepts that are strongly rooted in Cuban and Latin American society in general. This corresponds to a process of recruitment based on moral values. In this context Castro speaks in the name of socialist and Marxist ideals themselves, employing the *Me-Spokesman* discursive identity. Interestingly, some terms transcribing the favorable reaction of the crowd, are connected to these terms (e.g. 'aplausos' [applauses]), and illustrate the strength of such discursive strategies.

Defense of the project: This DP-discourse contains many verbs in infinitive with an imperative value (e.g. 'pelear' [to fight], 'defender' [to defend]), which call upon the audience to join and protect the plan, providing Castro with an Ethos of authority and power. In addition, we find some belligerent nouns (e.g. 'frente' [front], 'firme' [firm]) as well as combinations of words with a strong semantic impact, depicting disastrous scenarios (e.g. 'miseria' [misery], 'pobreza' [poverty], 'sangre' [blood]). These elements correspond to a discursive strategy of rhetoric of effects, since they seek to make the audience feel outraged and fearful, and to trigger a defensive or aggressive reaction in the crowd. According to Charaudeau, these types of strategies are common in political discourse, since an audience immersed in these feelings, would accept a message more easily. Using these elements, Castro summons the audience using the traditional triadic scenario: Marxist revolution is leading Cuban society towards a better future; however, there is an enemy that threatens the people and their future; so Castro calls on the audience to join and defend the project. This approach also corresponds to a process of recruitment, based on the defense of social, political, moral and national ideals.

Identification: The presence of the first person plural in verbs and possessive adjectives (e.g. 'nosotros' [we], 'nuestro' [our]) shows that Fidel Castro uses the *Me-Us* discursive identity. This strategy has a two-fold goal: It allows Castro to create a higher degree of identification with his audience, and it makes people perceive themselves as major actors in Castro's plan. Since Castro presents himself as a committed ruler, identified with the people, and connected to the beneficiaries of his policies, he exhibits an Ethos of authority, identification and charisma.

Veracity: In this DP-discourse we have identified different terms showing that Castro aimed at increasing the veracity of his discourse. First, there are several words denoting negation (e.g. 'nunca' [never], 'ninguna' [none]) that are connected to terms denoting the concept of absoluteness. Since these terms are also strongly connected to the term 'duda' [doubt] we deduced that Castro asked his audience to believe in him beyond any possible doubt. Hence, in order to convey the absolute veracity of his speeches, Castro exhibits an Ethos of power and authority. Moreover, in this context, there are also different terms denoting the first person singular (e.g. 'digo' [I say], 'creo' [I believe]). This suggests that Castro exhibits the *Me* discursive identity, expressing a personal commitment, and presenting himself as the guarantor of the veracity of his speech.

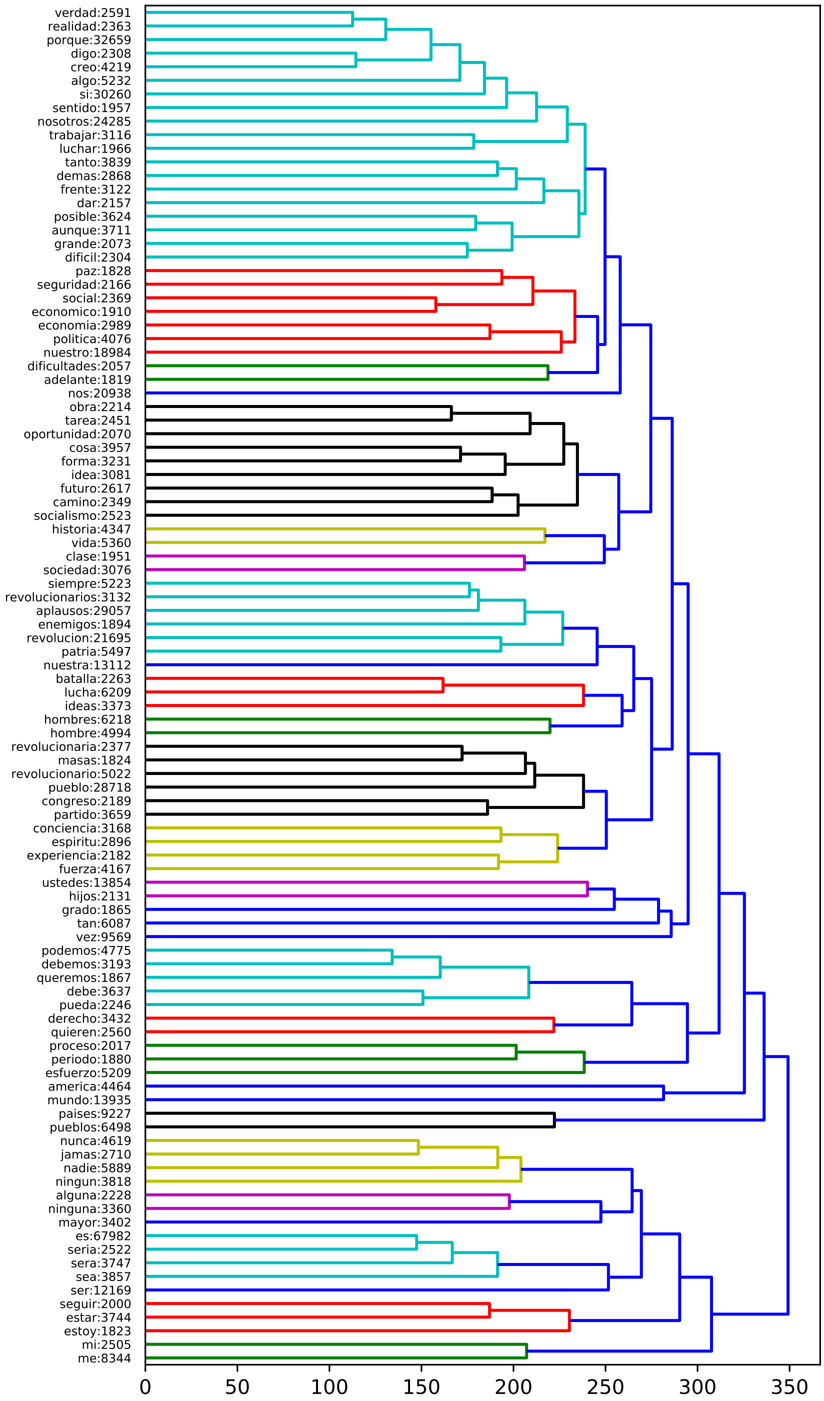

Figure 3: DP-discourse 466 containing the 100 most frequent terms.

Finally, there are several logical operators (e.g. 'porque' [because], 'aunque' [although], 'si' [if], 'tampoco' [nor]). Since these terms participate in argumentative processes by definition, we deduced that Castro aimed to increase the credibility and truthfulness of his speeches by using arguments belonging to the Logos family (possibly relying on simplified augmentation and self-evident assumptions).

Cuba's economic development (DP-discourse 93) This DP-discourse has a vocabulary of 881 words and 614,923 occurrences. The percentage of vocabulary size and the number of occurrences expressed in this DP-discourse are equal to 13.65% and 14.78% respectively. Analyzing this DP-discourse, we can infer that Castro's speeches often evoke Cuba's economic development. This concept is conveyed by several nouns from the lexical field of industrial and agricultural production, as well as verbs of necessity, obligation and action. In this context, we can also deduce that Castro mainly uses a rhetoric of effects and aims at increasing the legitimacy and the veracity of his message through detailed descriptions. A DP-discourse representing the 100 most frequent terms from cluster 93 is illustrated in Figure 4.

Needs, obligations, actions: This DP-discourse is characterized by the presence of several nouns and verbs in the present and future tense denoting needs and possibility (e.g. 'necesitan' [they need], 'pueden' [they can], 'tienen' [they have to]). Since these terms are linked to words referring to industrial and agricultural projects (e.g. 'industria' [industry], 'ganadería' [livestock], 'caña' [sugar cane]), we infer that Castro shows that these plans respond to the country's fundamental needs and open new opportunities for Cuba. This discursive strategy corresponds to a recruitment process, based on ideals of progress. Moreover, since this strategy also aims at creating in the audience feelings of hope and national pride, it also corresponds to a rhetoric of effects. We find several verbs from the lexical field of realization, in present and future tense (e.g. 'tendremos' [we will have], 'alcanza' [it reaches]). Fidel Castro is likely to have used these verbs to show that the projects were being executed and would be completed soon. Moreover, there are verbs of realization in subjunctive, which show that these plans were an important aspiration for Castro. There are also several terms from the lexical field of time that were probably used by Castro to refer to the execution schedule of the projects (e.g. 'diarias' [daily], 'mensuales' [monthly]). Castro provides these details in order to increase the veracity of his message and to show himself as an expert who masters all the details related to the execution of the projects. Moreover, using these mechanisms, Fidel Castro increases his own legitimacy, embodying an Ethos of commitment. Finally, the different verbs appearing in this DP-discourse are mainly conjugated in the first person plural and in the third person plural and singular (e.g. 'disponemos' [we dispose], 'crecen' [they grow]). The presence of the first person plural indicates that Fidel Castro uses the *Me-Us* discursive identity, showing his closeness and identification with the Cuban people.

Wealth and economic prosperity: This DP-discourse is characterized by the presence of many terms referring to the development of agricultural and industrial means of production (e.g. 'tractores' [tractors], 'fábrica' [factory]). Moreover, there are terms from the lexical field of wealth and production in general (e.g. 'bienes' [goods], 'riquezas' [resources], 'producto' [product]). We also find verbs expressing actions related to economy and production. And we mainly identified words corresponding to different economic resources and products, such as raw materials (e.g. 'caña' [sugar cane], 'petroleo' [oil]). These words are strongly linked to several terms conveying the concepts of quantity, quality and value (e.g. 'dólares' [dollars], 'toneladas' [tones], 'millones' [millions]). This suggests that Castro describes these projects in detail in order to increase the veracity and credibility of his speeches and also to show himself an expert on the subject. Moreover, this rich vocabulary creates an impression of wealth and economic prosperity. This strategy aims at generating in his audience a feeling of hope, well-being, security and pride, and thus corresponds to the rhetoric-of-effects strategy. On the other hand, since this strategy is based on the imaginary of progress

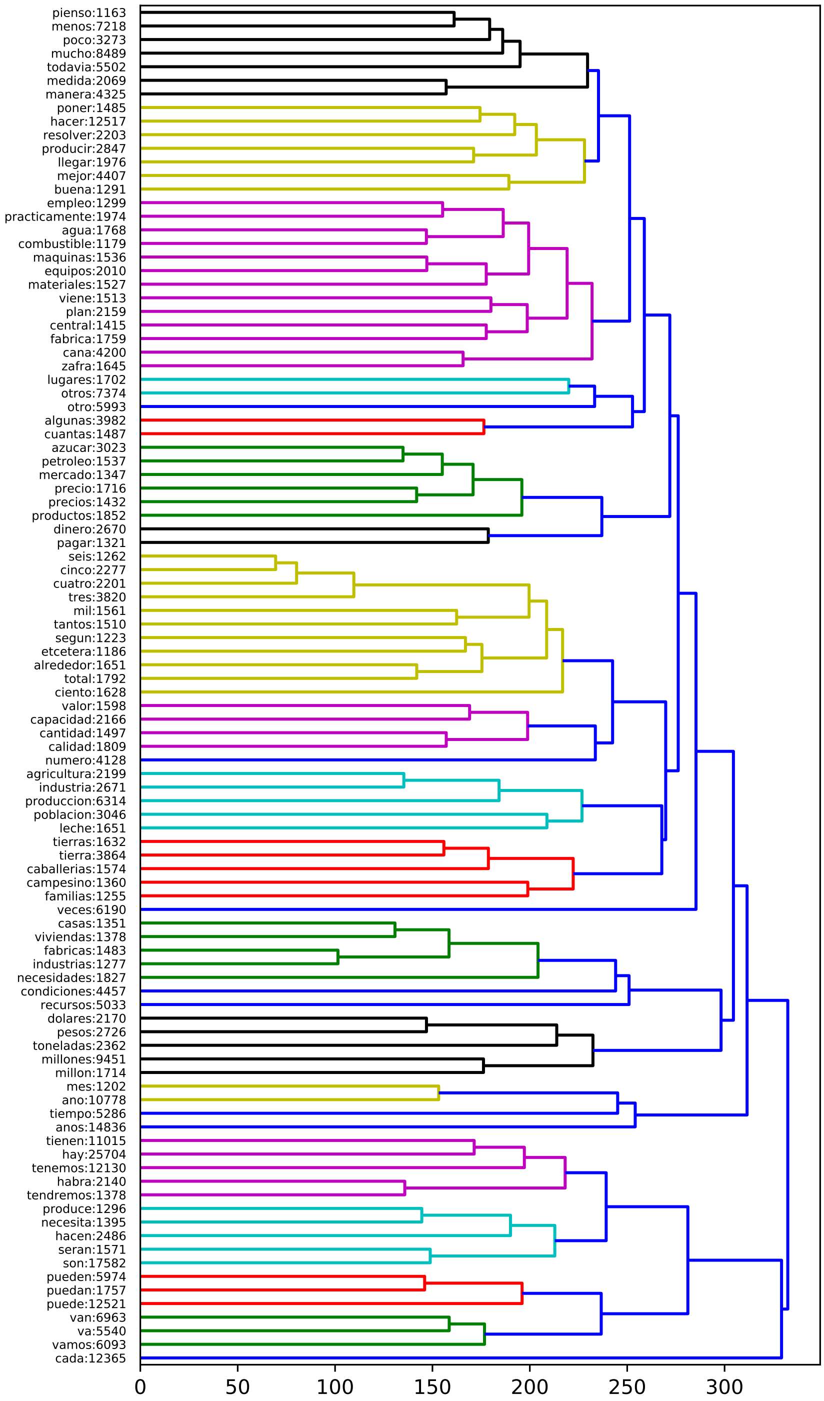

Figure 4: DP-discourse 93 containing the 100 most frequent terms.

and economic growth, it also corresponds to a recruitment process.

Social infrastructure This DP-discourse also contains several nouns that refer to the development of social infrastructure and services, such as public housing (e.g. 'casas' [houses], 'comida' [food], 'tienda' [store]). We also find nouns that evoke the beneficiaries of these social projects (e.g. 'familias' [families], 'pobres' [poor], 'campesino' [peasant]). This strategy corresponds to a recruitment process, based on the social welfare imaginary. Moreover, since this mechanism aims at generating a feeling of hope in the audience, it also corresponds to a rhetoric of effects.

Social welfare projects (DP-discourse 70) This DP-discourse has a vocabulary of 550 words and 476,957 occurrences. The percentage of vocabulary size and the number of occurrences expressed in this DP-discourse are equal to 8.52% and 11.46% respectively. In this DP-discourse, Fidel Castro presents the work that his government and himself are carrying out. These ongoing projects address social welfare issues and are mainly related to education, health and employment. A DP-discourse representing the 100 most frequent terms from cluster 70 is illustrated in Figure 5.

Ongoing projects: This DP-discourse contains lexicon referring to the development of projects (e.g. 'construyendo' [building], 'proyectos' [projects]). There are also verbs in the present tense conveying the idea of ongoing actions (e.g. 'creando' [creating], 'convirtiendo' [transforming], 'resolviendo' [solving]). These verbs are likely to increase the veracity of Castro's message by referring to tangible, ongoing steps. The aforementioned terms are linked to adjectives and adverbs that describe these steps using a range of positive connotations such as importance, variety, novelty and quantity (e.g. 'importantes' [important], 'enorme' [huge], 'mejores' [best], 'diversos' [diverse]). These elements aim at captivating the audience by creating feelings of hope and pride; they thus correspond to a rhetoric of effects. Furthermore, Castro evokes the places where such projects are being executed (e.g. 'ciudad' [city], 'región' [region], 'Camaguey', 'Matanzas', 'Cienfuegos'). These geographical details increase the veracity and the credibility of the speeches and, in turn, indicate that the entire country benefits from the projects. This corresponds to a recruitment process by national belonging.

Society as an actor: Among the terms referring to ongoing projects, there are several infinitive verbs (e.g. 'realizar' [to make], 'participar' [to participate]), and words referring to people (e.g. 'obreros' [worker], 'trabajadores' [workers], 'sindicatos' [unions]). Since verbs in the infinitive form may have a value of imperative, we can infer that Castro calls on the audience to participate in these projects, as individuals or as members of social organizations. Therefore, Castro shows himself as the organizer, and the head of the projects, exhibiting an Ethos of authority and commitment. In addition, Castro uses possessive adjectives in the first person plural (e.g. 'nuestros' [ours]). This shows that Castro uses the *Me-Us* discursive identity in order to include society in the development of the projects. This also indicates that Castro develops an identification process between him and the Cuban people.

Education, employment and health: Specific vocabulary suggests that the projects described by Castro embrace three major social welfare objectives: education (e.g. 'estudiantes' [students]), employment (e.g. 'empleos' [employment]) and health (e.g. 'salud' [health]). These elements are major social welfare topics and they directly imply a better quality of life for the population, which is shown as the direct beneficiary of the projects. The objective of this strategy is to generate feelings of hope and well-being in the audience; this corresponds to a rhetoric-of-effects strategy. In addition, the direct reference to ideals of social welfare induce a recruitment process, since society is likely to share such ideals. Finally, the large amount of specific lexicon in this DP-discourse contributes to the veracity of the speeches and allows Castro to present himself as an expert.

Cold War and triadic scenario (DP-discourse 344) This DP-discourse has a vocabulary of 989 words and 457,130 occurrences. The percentage of vocabulary size and

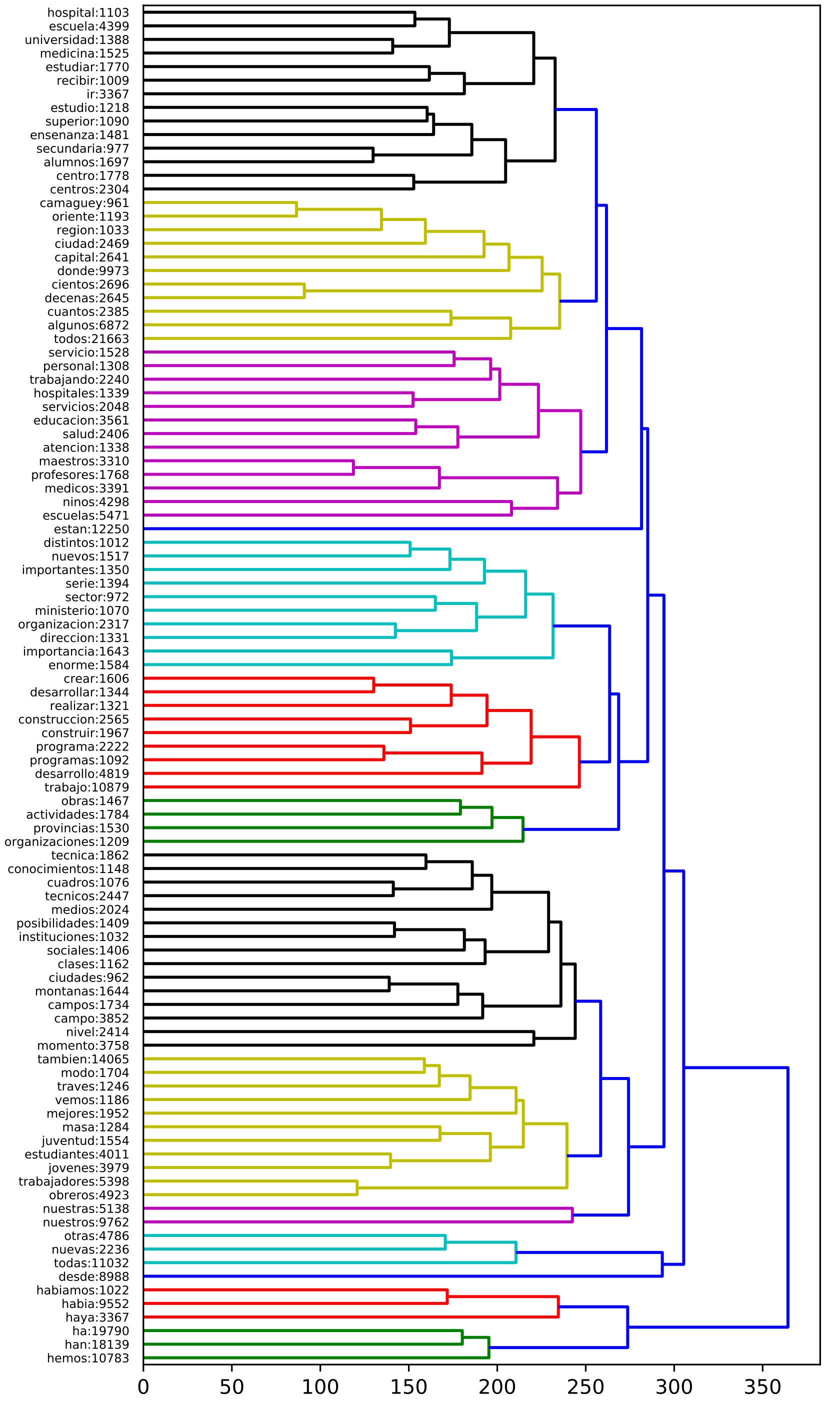

Figure 5: DP-discourse 70 containing the 100 most frequent terms.

the number of occurrences expressed in this DP-discourse are equal to 15.33% and 10.98% respectively. Here, Castro reports different historical conflicts that occur in Cuba and around the world, mainly in the context of the Cold War. These events exhibit a particularly disastrous situation and correspond to the first element of the triadic scenario. In this context, Castro characterizes the main enemies of the Cuban people: the United States and local elites. This description corresponds to the second element of the triadic scenario. Finally, the third element corresponds to the socialist Bloc, communist parties, people and the army of Cuba. A DP-discourse representing the 100 most frequent terms from cluster 344 is illustrated in Figure 6.

Historical conflicts and wars: This DP-discourse evokes different historical conflicts that took place in the context of the Cold War in Cuba and around the world: the dictatorial government of Fulgencio Batista overthrown by the Cuban revolution (e.g. 'Batista', 'dictadura', 'Moncada'); guerrilla movements (e.g. 'guerrilla'); the attempted invasion of Playa 'Girón' (Bay of Pigs Invasion) by the United States; the economic 'embargo' promoted by the United States against Cuba; the civil war in 'Angola'; the 'Vietnam' War; Latin American dictatorships ('Trujillo' [Dominican Republic], 'Somoza' [Nicaragua], 'dictador' [dictator]); the wars of decolonization (e.g. 'Argelia', 'Guinea'). These elements are connected to time markers (e.g. 'julio' [July], 'abril' [April]), and verbs in the past tense (e.g. 'hubo' [there was], 'hicieron' [they did]), showing that Castro includes analogies with the past in his speeches. This allows him to increase the legitimacy and the veracity of his discourse and depict himself as an expert on history and international politics. Furthermore, the description of the historical conjuncture is strongly impregnated by military lexicon (e.g. 'infantería' [infantry], 'artillería' [artillery]), and moral and national values ('soberanía' [sovereignty], 'democracia' [democracy], 'libertades' [freedoms]). References to war and ideals create feelings of fear, heroism, and pride in the audience, which corresponds to the rhetoric of effects. This strategy also corresponds to a recruitment process based on nationalism and moral values. Finally, the details conveyed by the military vocabulary aim at increasing the veracity of the discourse.

Cold War and triadic scenario: Castro's description of historical events is defined by the Cold War and its focus on conflicts between the socialist and the western Blocs. These Blocs constitute the binary conception of the world depicted by Castro and they are organized following the triadic scenario, where the western and the socialist Blocs correspond to the *Enemy* and the *Hero* respectively. This binary description is a simplification of a complex scenario. This presentation allows the audience to concentrate on a few ideas and thus corresponds to the singularization argument. This mechanism is often complemented by the argument of essentialization, which condenses complex ideas into a few simple terms. Hence, we deduce that some terms (e.g. 'imperio' [empire], 'burgués' [bourgeois]) may correspond to this mechanism. The combination of singularization and essentialization, correspond to the use of argumentative formulas that generate a strong feeling of evidence and attraction.

Enemy: This DP-discourse is characterized by vocabulary evoking the existence of an external enemy, which is depicted as the opponent to liberation movements around the world. The enemy is clearly associated with the United States, using different nouns (e.g. United States, 'imperio' [empire], 'Yankis' [Yankees]). Castro also describes an internal enemy, represented by the economically dominant class (e.g. 'burgués' [bourgeois]), and by the political elites (e.g. 'oligarquía' [oligarchy], 'politiqueros' [demagogues]), which receive orders from the external enemy (e.g. 'esbirros' [henchmen], 'títeres' [puppets]). The speaker uses a Marxist conception of society to link the economic elites to the concept of labour exploitation (e.g. 'explotadores' [exploiting]). On the other hand, Castro associates the local elites, and especially the external enemy, with crimes and destruction (e.g. 'genocida' [genocidal], 'destrucción' [destruction]); these terms create an extremely negative image of the enemy. These elements contribute to depict the enemy as numerous and dangerous and create the feeling in the audience that there is a hidden threat. In this context, Castro implicitly depicts a latently disastrous situation. The goal of this rhetoric of effects is to generate fear and anxiety in the audience, in order to predispose people to accept his message.

Hero: The hero described by Fidel Castro has several facets, which he most probably adapted according to the audience and to the political context. First, Castro assigns a major role to the Cuban revolutionary army and rebel groups: different nouns present this institution as the vanguard in the fight against the enemy ('guerrilleros' [guerrilla groups], 'milicianos' [militiamen], 'cubanos' [Cubans]). Moreover, there is a large lexicon evoking civil society and socialist political movements related to Castro's government (e.g. 'socialista' [socialist], 'comunista' [communist], 'gente' [people]). Therefore, we deduce that Fidel Castro creates an amalgam between communist organizations, the people and the army of Cuba, producing a greater sense of identification between and among these actors. Furthermore, the description of these actors is associated to values and ideals, using nouns and adjective with strong semantic impact (e.g. 'heroismo' [heroism], 'sacrificio' [sacrifice]). This strategy aims at creating feelings and passions in the audience and thus corresponds to a rhetoric of effects. In addition, by using values and ideals as banners in the fight against the enemy, Fidel Castro also applies a recruitment process, based on moral, social, national and ideological belonging. Finally, Castro evokes different important historical characters, such as José Martí, the Cuban writer, politician and hero of the Cuban independence war and the well-known guerrilla commander Ernesto Che Guevara. Castro evokes José Martí, to create nationalist feelings in his audience and also to increase his legitimacy by presenting himself as the successor of this historical figure, while Che Guevara is invoked as an heroic example of a fighter against the enemy. Castro thus creates a revolutionary pantheon of heroes.

Summary In order to have an overview of Castro's discursive strategies, we decided to count the number of times each strategy was found in the DP-discourses. To do so, we conducted an exhaustive manual expert-driven analysis of the discursive strategies present in the branches of the four most important DP-discourses. For each discursive strategy that was detected in a branch, twenty random sentences from the corpus, containing the involved words, were extracted and systematically checked in order to confirm our interpretations. According to this study, Pathos, Ethos and Logos make up 51%, 34% and 15% of these strategies respectively, which is coherent with general characteristics of political speeches as described by Charaudeau (2005). The strategies' frequencies are illustrated in Figure 7. According to this radar-chart, the most frequently used Pathos strategies are rhetoric of effects, recruitment process based on social ideals, nationalism and progress, and the triadic scenario. The most frequently used Ethos strategies are the Me-Us discursive identity, the Ethos of authority, commitment, expertise and the identification with the audience. Finally, the veracity strategy seems to be the most frequently used of Logos strategies.

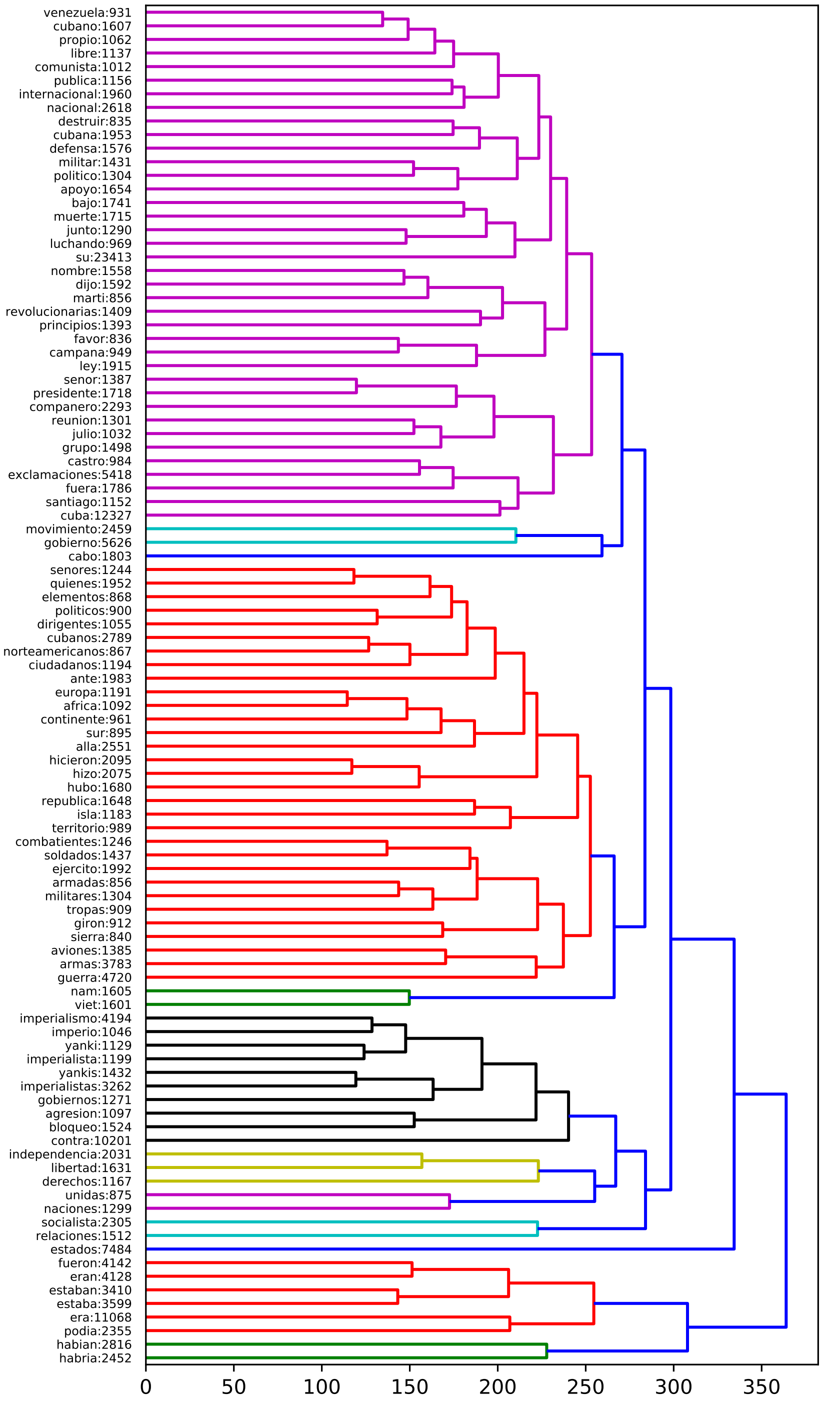

Figure 6: DP-discourse 344 containing the 100 most frequent terms.

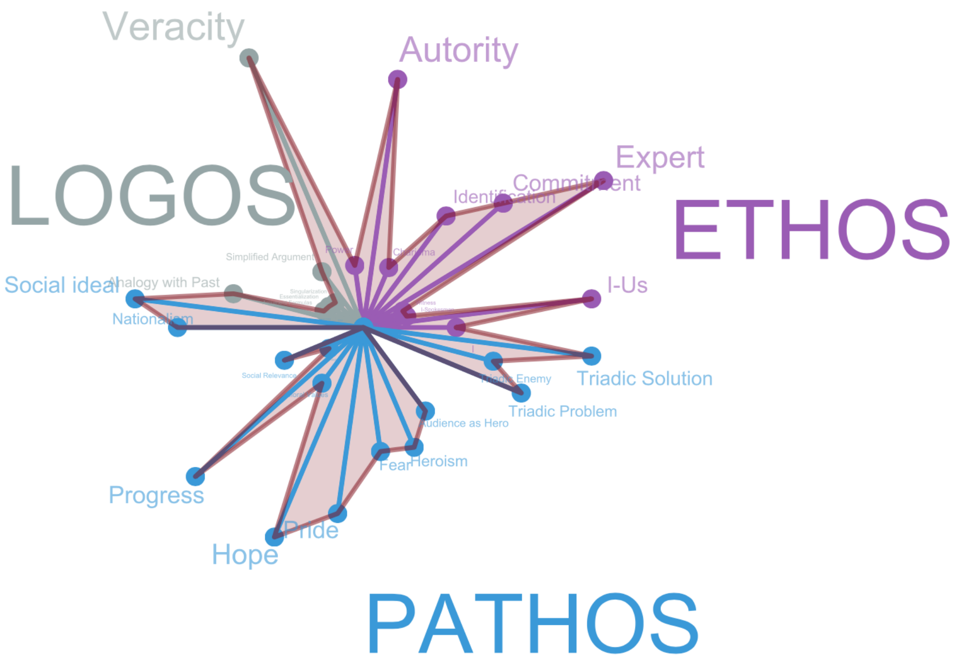

Figure 7: Radar-chart summarizing the discursive strategies of Fidel Castro. Each radius represents one strategy, and its length is proportional to its frequency across the full analysis. The Pathos, Ethos and Logos strategies are respectively colored in blue, violet and green.

Reference	**Methodology**	**Discursive strategies**	**Number of strategies**
Joyner (1964)	Aristotelian rhetoric	People-hero, people-victim, enemy, moral feelings: anger, fear, confidence, hope committed, expert, identification causal, analogy, induction, deduction	12
Fagen (1965)	Non-lexicometric	Castro-hero, people-victim, enemy authority, identification, charisma	6
Nieto et al. (2002)	Conversational analysis	Charisma, identification	2
Belisario (2010)	Cognitive linguistics Critical disc. analysis	People-hero, people-victim, enemy authority, identification	5
Reyes (2011)	Sociolinguistics Critical disc. analysis	Witness, expert, charisma, identification veracity	5
De Sousa (2009a) De Sousa (2012) De Sousa (2009b)	Lexicometric	People-hero, people-victim, enemy moral, ideology identification	6
DP-discourses	Hybrid: Data mining - Semio-pragmatic	People-hero, Castro-hero, people-victim enemy, nationalism, ideology progress, welfare, moral feelings: anger, fear, confidence, hope committed, witness, expert, authority charisma, identification causal, analogy, veracity	**19**

Table 2: Discursive strategies reported previously in the literature, and in this work using DP-discourses (last entry). For the sake of clarity, strategies have been coloured according to their category, following the Aristotelian classification of rhetoric. Ethos, Pathos and Logos are reported in violet, blue and green, respectively.

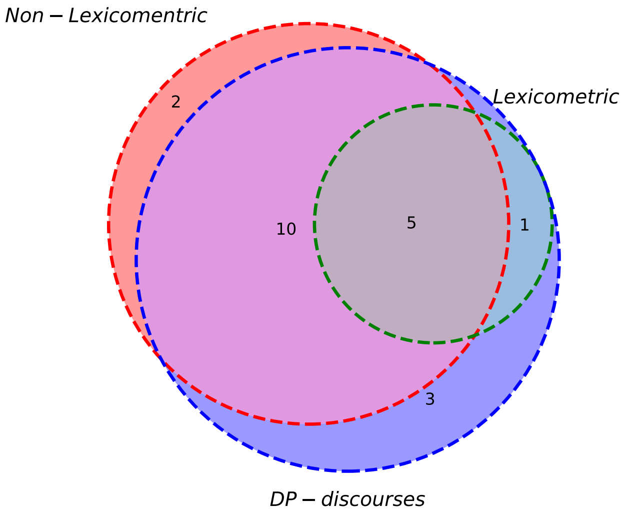

Figure 8: Venn diagram representing the number of discursive strategies found by lexicometric approaches (green circle), non-lexicometric approaches (red circle) and those retrieved using DP-discourses (blue circle).

7 Discussion

In this work, we hypothesized that the use of a data mining driven methodology could provide a more representative characterization of Castro's discursive strategies. In order to assess this hypothesis, and to illustrate the contribution of the paper to the existing literature, we compared our conclusions with those of previous works. The discursive strategies reported so far in the literature are summarized in Table 2, and the number of discursive strategies found by general families of methods (i.e. lexicometric, non-lexicometric and based on DP-discourses) are represented using a Venn diagram in Figure 8.

As shown in Table 2 most of the previous works, focused on particular aspects of Castro's discursive strategies: Nieto et al. (2002) and Reyes (2011) mostly studied the discursive identity of Castro, while Belisario (2010) and Fagen (1965) analyzed Castro's triadic scenario and his discursive identity. De Sousa (2009a,b, 2012) focused on the evolution of the major topics developed in the speeches, with respect to the historical scenario. Finally, Joyner (1964) was the only previous work that aimed at presenting a general picture of Castro's discursive strategies, using a non-lexicometric methodology derived from Aristotelian rhetoric.

As depicted in Table 2 and Figure 8, this work reports the highest number of discursive strategies, i.e. 19 in total, including 16 out of 18 strategies that were described by at least one of the eight previous works.

The two missing strategies, namely inductive and deductive reasoning, have only been reported by Joyner (1964). These strategies belong to the Logos family, and they are usually carried by complex articulations of words, which are likely to be lost by Vector Space Modeling algorithms. Indeed, Vector Space Models capture the semantic relationship between words, by assigning similar vector representations to words appearing in the same context; as a result, fairly rare and complex associations of words are likely to be lost at the expense of more frequent ones. In this case, specialized approaches, such as a systematic expert-driven analysis, should be applied to retrieve these strategies.

On the other hand, three important discursive strategies that were found in this work have been neglected in the literature, possibly due to small corpus selection biases. Indeed, previous studies have not focused on the importance of recruitment processes based on ideals of progress, social welfare and nationalism. Nevertheless, as presented in Section 6.2, these discursive strategies have been widely used by Castro, with the objective to lead his audience to accept his project willingly.

Consequently our approach was able to draw a representative landscape of Castro's discursive strategies that agrees with the major conclusions of previous works, while also offering new insights into this research question.

8 Conclusion

Summary This paper presents a two-fold contribution to the digital humanities community: On the one hand, we propose a new discourse analysis data mining framework to study large data copus. This new approach combines state-of-the-art data mining tools with the well-known semio-pragmatic linguistic discourse analysis methodology On the other hand, we have provided a broad and representative characterization of the main discursive strategies used by Castro. This study was conducted on a large corpus of more than 1,018 speeches and 7,500,000 words, combining state-of-the-art data mining tools and the semio-pragmatic discourse analysis methodology. According to this study, Castro presents himself as an authority, an expert committed to his duties and identified with his audience. His speeches are organized around a Cold War triadic scenario, his government and socialist movements worldwide are presented as heroes that protect people against the source of all problems, i.e. the enemy which is represented by the USA and the bourgeoisie. In this context, he shows the audience as a potential hero and beneficiary, and he alludes to the progress made by his country. These elements evoke strong feelings such as heroism, pride, hope and fear. Finally, Castro tends to include many details to increase the veracity of his speeches. A comparison between these findings and those of previous works reveals that our method confirmed most of the previously reported discursive strategies and provided new insight into the importance of recruitment processes based on nationalism, welfare and progress. These discursive strategies aim at captivating the audience by incorporating references to classic positive values such as social welfare policies, ideals of technological progress, economic growth and nationalist ideals. Thus, our study provides a broad and representative characterization of the main discursive strategies used by Finally Castro.

Wider relevance Given that our methodology resulted in a representative characterization of Castro's rhetoric strategies, it seems promising to apply it to other case studies, and to use it to analyze political rhetoric in a broader context. For instance, considering the resurgence of populism in Europe and America (Berezin, 2009, De la Torre, 2010, Greven, 2016), it could be particularly interesting to characterize the rhetorical strategies of populist leaders. Indeed, according to Jansen (2011), populist movements are mainly based on two components: popular mobilization and populist rhetoric. Characterizing, and analyzing, the discursive strategies of various populist leaders is therefore an important step towards understanding this phenomenon. In this context, a hybrid discourse analysis methodology, as presented in this paper, could provide a representative and broad overview of the populist rhetoric and make it possible to unravel common populist discursive strategies. Furthermore, this kind of framework could facilitate the elaboration of fast, large-scale and more objective analyses of political discourses at critical moments (i.e. elections, political crisis), which could have a direct impact on society. Interestingly, this research topic has recently motivated the creation of global research projects such as the Populismus Observatory, [5] and the Team Populism. [6] Both projects aim at developing research communities to promote the study of populist rhetoric by sharing data, information and tools. In this context, the framework presented in this paper could be integrated into such projects as a complementary analysis tool.

Perspectives and Future Work In the future, we plan into incorporate the temporal dimension to the discourse analysis. The corpus of Fidel Castro's speeches is particularly well adapted to these kinds of temporal considerations, since the date of each speech is also reported. Simply by counting the number of words belonging to each DP-discourse at different periods of time and by following the frequency changes, we could study the evolution of Castro's discursive strategy over time and with respect to the speeches' historical context. Another promising research path consists in using our data mining framework to characterize the discursive strategies of other politicians, and possibly using other non-lexicometric discourse analysis methodologies. Future perspectives also include a thorough assessment of this method and a comparison with alternative approaches that could rely on different pre-processing steps (e.g. Brants (2000)), different word embedding techniques (e.g. Le and Mikolov (2014)), alternative subspace clustering methods (e.g. Kriegel et al. (2009)), topic modelling algorithms (e.g. Blei et al. (2003), Jain et al. (1999)) and complementary visualization techniques (e.g. Maaten and Hinton (2008)).

References

Agarwal, S. and H. Yu
2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Aggarwal, C. C., A. Hinneburg, and D. A. Keim
2001. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, Pp. 420–434. Springer.

Aggarwal, C. C., J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park
1999. Fast algorithms for projected clustering. In *ACM SIGMoD Record*, volume 28, Pp. 61–72. ACM.

Barry, A.
2002. Les bases théoriques en analyse du discours. *Documents de la Chaire MCD*, 159.

Belisario, A. G. V.
2010. Sistemas metafóricos en discursos de Fidel Castro: "decir la verdad en el primer deber de todo revolucionario". *Letras*, (81):139–162.

Benzécri, J.-P. et al.
1973. *L'analyse des données*, volume 2. Dunod Paris.

[5] http://observatory.populismus.gr/

[6] https://populism.byu.edu

Berezin, M.
2009. *Illiberal politics in neoliberal times: culture, security and populism in the new Europe*. Cambridge University Press.

Bird, S., E. Klein, and E. Loper
2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Blei, D. M., A. Y. Ng, and M. I. Jordan
2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brants, T.
2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, Pp. 224–231. Association for Computational Linguistics.

Caliński, T. and J. Harabasz
1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Charaudeau, P.
1995. Une analyse sémiolinguistique du discours. *Langages*, Pp. 96–111.

Charaudeau, P.
2005. Quand l'argumentation n'est que visée persuasive. l'exemple du discours politique. *Argumentation et communication dans les médias. Québec: Éditions Nota Bene*, Pp. 23–43.

Charaudeau, P.
2008. Pathos et discours politique. *Émotions et discours. L'usage des passions dans la langue. Rennes: Presses universitaires de Rennes*, Pp. 49–58.

Charaudeau, P.
2009a. Identité sociale et identité discursive. un jeu de miroir fondateur de l'activité langagière. *Identités sociales et discursives du sujet parlant*, Pp. 15–18.

Charaudeau, P.
2009b. Le discours de manipulation entre persuasion et influence sociale. In *Acte du colloque de Lyon*.

Charaudeau, P.
2011. Réflexions pour l'analyse du discours populiste. *Mots. Les langages du politique*, (97):101–116.

Church, K. W. and P. Hanks
1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

De la Torre, C.
2010. *Populist Seduction in Latin America*. Ohio University Press.

De Sousa, S.
2009a. Le discours de Fidel Castro. Éssai de lexicométrie politique. *Lexicometrica, Explorations textométriques*, 2:68–94.

De Sousa, S.
2009b. Le peuple dans le discours Fidel Castro. *Communication au Colloque Représentation du Peuple*, Pp. 1–14.

De Sousa, S.
2012. À l'épreuve des temps... temps lexical et temps politique dans le discours de Fidel Castro (1959-2008). *A. Dister, D. Longré et G. Purnelle (Éds), JADT*, Pp. 337–349.

Ekman, P.
1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ellis, P.
2010. The essential guide to effect sizes: An introduction to statistical power. *Meta-Analysis and the Interpretation of Research Results.: Cambridge University Press.*

Fagen, R. R.
1965. Charismatic authority and the leadership of Fidel Castro. *Western Political Quarterly*, 18(2-1):275–284.

Fairclough, N.
2013. *Critical discourse analysis: The critical study of language*. Routledge.

Gee, J. P.
2014. *An introduction to discourse analysis: Theory and method*. Routledge.

Gott, R.
2007. *Cuba*. Ediciones Akal.

Greven, T.
2016. The rise of right-wing populism in europe and the united states. *A Comparative Perspective. Friedrich Ebert Foundation, Washington DC Office.*

Harris, Z. S.
1954. Distributional structure. *Word*, 10(2-3):146–162.

Hudson, R. A.
1996. *Sociolinguistics*. Cambridge university press.

Jain, A. K., M. N. Murty, and P. J. Flynn
1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Janis, I. L. and L. Mann
1977. *Decision making: A psychological analysis of conflict, choice, and commitment.* Free press.

Jansen, R. S.
2011. Populist mobilization: A new theoretical approach to populism. *Sociological theory*, 29(2):75–96.

Jones, E., T. Oliphant, P. Peterson, et al.
2019. SciPy: Open source scientific tools for Python.

Joyner, G. M.
1964. *Persuasive elements in the speeches of Fidel Castro*. PhD thesis, Texas Tech University.

Kriegel, H.-P., P. Kröger, and A. Zimek
2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1):1.

Le, Q. and T. Mikolov
2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, Pp. 1188–1196.

Lövheim, H.
2012. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical hypotheses*, 78(2):341–348.

Luong, X. and S. Mellet
2003. Mesures de distance grammaticale entre les textes. *Corpus*, (2).

Maaten, L. v. d. and G. Hinton
2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Maingueneau, D.
2016. *Les termes clés de l'analyse du discours*. Le seuil.

McDonald, S. and M. Ramscar
2001. Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

Mikolov, T., K. Chen, G. Corrado, and J. Dean
2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean
2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, Pp. 3111–3119.

Nieto, M. J. et al.
2002. La afectividad en la comunicación política. *Opción: Revista de Ciencias Humanas y Sociales*, (39):36–53.

Padmos, R. et al.
2017. Fidel and raúl castro's ideological influence on foreign policy in reaction to the us trade embargo. B.S. thesis.

Pêcheux, M.
1995. *Automatic discourse analysis*, volume 5. Rodopi.

Peignier, S., C. Rigotti, A. Rossi, and G. Beslon
2018. Weight-based search to find clusters around medians in subspaces. In *Proceedings of the ACM Symposium on Applied Computing*. ACM.

Plutchik, R.
2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Ramonet, I.
2010. *Fidel Castro: biografía a dos voces*. Debate.

Rehurek, R. and P. Sojka
2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Reyes, A.
2011. *Voice in political discourse: Castro, Chavez, Bush and their strategic use of language*. A&C Black.

Richards, J. C. and R. W. Schmidt
1983. Conversational analysis. *Language and communication*, Pp. 117–154.

Richardson, L.
2019. Beautiful soup.

Rubenstein, H. and J. B. Goodenough
1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Schwarz, N.
2000. Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440.

Sokal, R. R.
1958. A statistical method for evaluating systematic relationship. *University of Kansas science bulletin*, 28:1409–1438.

Turney, P. D. and P. Pantel
2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Van Rossum, G. and F. L. Drake
1995. *Python library reference*. Centrum voor Wiskunde en Informatica.

Weizman, E.
2008. *Positioning in media dialogue: Negotiating roles in the news interview*, volume 3. John Benjamins Publishing.

Widdowson, H. G.
1995. Discourse analysis: a critical view. *Language and literature*, 4(3):157–172.

Wiedemann, G.
2013. Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research/Historische Sozialforschung*, Pp. 332–357.