What can Reveal 1,018 Speeches of Fidel Castro?

Sergio Peignier, Patricia Zapata

What can Reveal 1,018 Speeches of Fidel Castro?

Sergio Peignier[+] **Patricia Zapata**[*]

[+] CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France

[*] Carrera de Lingüística e Idiomas, Universidad Mayor de San Andrés, La Paz, Bolivia

[+] Sergio.Peignier.Zapata@gmail.com

Introduction

Fidel Castro, the leader of the Cuban Revolution, has influenced different left-wing regimes and political movements in Latin America, and around the World. His ability to seduce masses relied largely on his rhetorical abilities. Indeed, according to Patrick Charaudeau, speech skills are crucial for a politician to captivate his audience to join his cause. Therefore, studying Castro's political speech is a crucial step towards understanding his political success. This problem has been addressed previously (e.g., [1, 2]). These studies were achieved on small discursive samples, with only few speeches. However, using a small and possibly non-representative sample is likely to lead to biased results. To avoid such problem, the work presented here was carried out on a large corpus of 1,018 speeches and more than 4,000,000 words, combining machine learning tools and the linguistic speech analysis methodology of P. Charaudeau. Using the association of both techniques, we provide here a more representative characterization of Castro's main discursive strategies.

Materials and Methods

Firstly, we proceeded to collect Castro's political speeches from the official website of the Cuban government (http://www.cuba.cu/gobierno/discursos/). Data collection was ensured by Python 2.7 web-scraping custom scripts relying on "urllib2" and "BeautifulSoup" libraries; whereas data cleaning and empty words filtering were ensured by Python 2.7 custom scripts using on the "re" and "nltk" libraries. In total, the corpus contains 1,018 speeches in spanish.

Secondly, we analyzed the corpus using Word2Vec [3], a well known Word Embedding algorithm based on neural networks. Word2Vec represents the words of a corpus as numeric vectors, building a so called Vector Space Model (VSM) of the Corpus. VSMs are able to capture the contextual and semantic relationship between words, so words that appear in the same context, tend to have similar vector representations. Word2Vec builds a VSM by feeding its neural network with words from a text, and training it to retrieve the words' neighbors within a given window range. Word2Vec has three main parameters, the size of the

vectors (here size=300), the neighborhood window size (here window=5), and the number of iterations over the corpus (here iterations=5). In this work we used the Word2Vec implementation available in "gensim" Python 2.7 library.

Once the VSM was built, we clustered its word vector representations. The aim of this step was to find groups of words that co-occur in the same discursive context, in order to analyze them separately. This task could have been achieved using any traditional clustering technique; however recent studies [4] have shown that traditional data mining algorithms struggle in high dimensional spaces, such as our 300 dimensional VSM. To overcome this problem, an alternative is to use subspace clustering instead. This 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 evolutionary subspace clustering algorithm called KymeroClust [5]. This technique is based on a K-medians paradigm, it groups data points around centers, and evolves the centers' coordinates and subspaces to minimize the distance to their closest data points. KymeroClust obtained competitive results compared to state-of-the-art algorithms, while it relies on a simple parameter setting procedure, and it adapts automatically the number of clusters.

Then, the words contained in the clusters were organized in Directed-Trees, using an algorithm inspired from [6]. These structures were built incrementally, using the VSM representations in order to reflect the contextual and semantic relationship between words.

These two last algorithms were implemented in Python 2.7.

Finally, Directed-Trees were studied using the speech analysis methodology developed by Patrick Charaudeau [7, 8]. This methodology identifies three families of discursive strategies: Ethos Pathos and Logos. The Ethos strategies allow the politician to build his discursive identity, they provide him with credibility, and they enable the audience to identify with the politician, which helps him to captivate the public. Pathos strategies address the most affective and emotional part of the audience, and their goal is to bring out feelings and passions. Finally, the Logos strategies use arguments to address the most rational part of the public.

Results and Conclusions

The application of the previous procedure generated 33 clusters; the seven largest ones contained 75.06% of the words, while the remaining groups included mostly outliers. Each one of these clusters encompasses only tightly related discursive contexts and topics; and their corresponding Directed-Trees revealed to be comprehensible. These observations are illustrated in Figures 1 and 2: Both figures correspond to samples of Directed-Trees obtained from different clusters. The first one contains terms alluding to the concept of Enemy: the USA designed by Castro as "Imperio [Empire]"; while other words allude to negative actions, e.g., "Agresión [Aggression]". The second sample is mostly related to economic aspects: "Petróleo [oil]", "Precios [prices]".

In order to characterize Castro's rhetoric, we counted the number of times each strategy was encountered in the Directed-Trees. These results provide a broad and representative characterization of Castro's speeches, as illustrated in the radar-chart Figure 3. We conclude this work summarizing Castro's rhetorical strategy:

Ethos: Castro exhibits himself as an authority, an expert committed to his duties and identified to his audience (he usually employs the inclusive first person of plural (I-Us) in his speeches).

Pathos: The so called "Triadic scenario" frames Castro speeches: The politician is a hero that protects people against an enemy, the source of all problems. In this context, he shows the audience as a potential hero and he alludes to the progress of his country. These elements evoke strong feelings such as heroism, pride, hope and fear.

Logos: Castro tends to include many details to increase the veracity of his speeches.

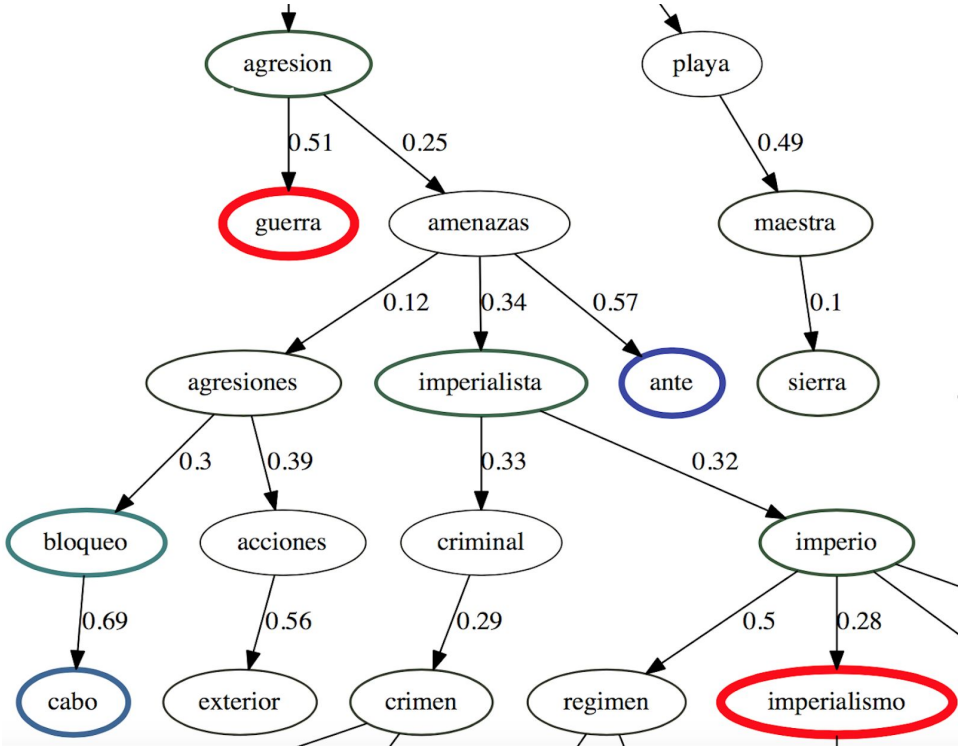

Figure 1: Directed-Tree sample evoking the Enemy ("imperio [empire]") and its actions ("agresion [aggression]", "bloqueo [embargo]").

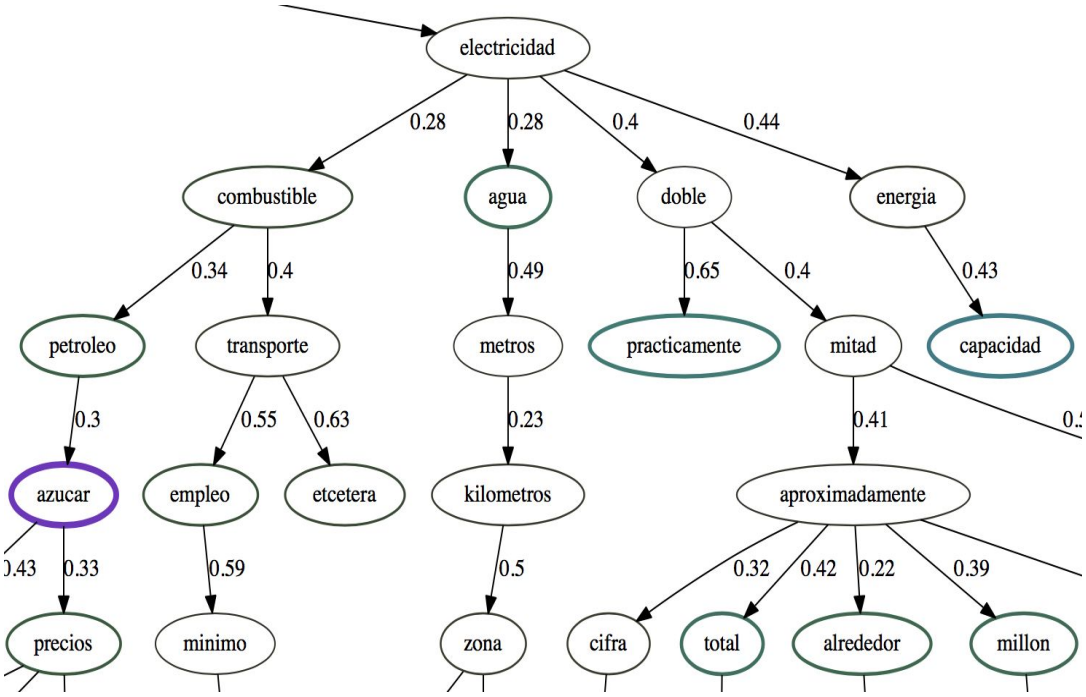

Figure 2: Directed-Tree sample evoking economic aspects ("petróleo [oil]", "azúcar [sugar]", "precios [prices]").

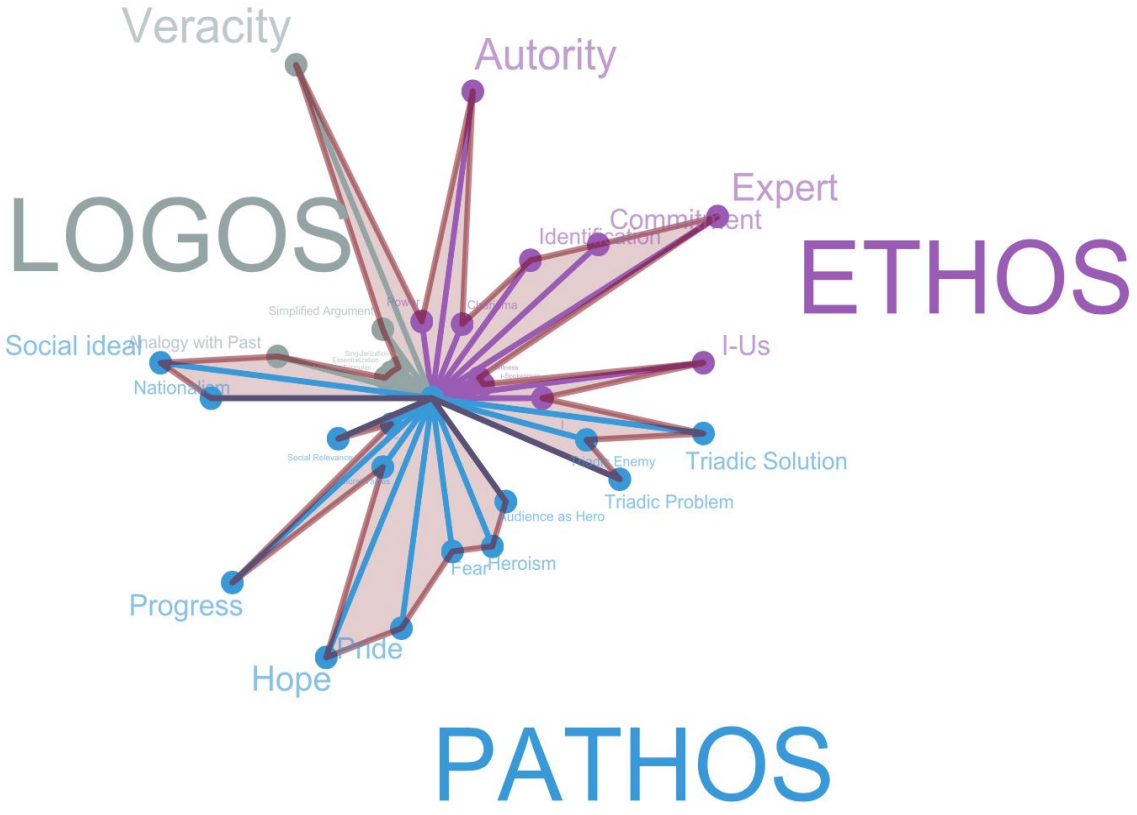

Figure 3: Radar-chart summarizing the discursive strategies of Fidel Castro.

Bibliography

[1] Bajini, I. (2010). Para una aproximación a la (r)evolución del discurso político latinoamericano desde Fidel Castro hasta Rafael Correa. Altre Modernità, (3), 133-155.

[2] León Guerra, F., Molero de Cabeza, L., & Chirinos, A. (2011). El discurso político en Latinoamérica. Análisis semántico-pragmático. Quórum Académico, 8(1).

[3] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[4] Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional spaces. In ICDT(Vol. 1, pp. 420-434).

[5] Peignier, S., Rigotti, C., & Beslon, G. (2017) EvoMove: Evolutionary-based living musical companion. European Conference on Artificial Life. p.8.

[6] Joglekar, S., (2015, October 15) Generating rudimentary Mind-Maps from Word2Vec models [Blog post]. Retrieved from https://codesachin.wordpress.com/2015/10/15/generating-rudimentary-mind-maps-from-word2vec-models/

[7] Charaudeau, P. (1993). Le contrat de communication dans la situation de classe. Inter-actions: l'interaction, actualités de la recherche et enjeux didactiques. Metz: Centre d'analyse syntaxique de l'Université de Metz, 121-136.

[8] Charaudeau, P. (2007). Les stéréotypes, c'est bien. Les imaginaires, c'est mieux. Stéréotypage, stéréotypes: fonctionnements ordinaires et mises en scène. Paris: L'Harmattan, 23-28.