Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 928–932
Marseille, 11–16 May 2020

A Corpus of Spanish Political Speeches from 1937 to 2019

Elena Álvarez-Mellado
Department of Computer Science, Brandeis University
415 South St, Waltham, MA 02453
ealvarezmellado@brandeis.edu

Abstract

This paper documents a corpus of political speeches in European Spanish. The documents in the corpus belong to the Christmas speeches that have been delivered yearly by the head of state of Spain since 1937. The historical period covered by these speeches ranges from the Spanish Civil War and the Francoist dictatorship up until today. As a result, the corpus reflects some of the most significant events and political changes in the recent history of Spain. Up until now, the speeches as a whole had not been collected into a single, systematic and reusable resource, as most of the texts were scattered among different sources. The paper describes: (1) the composition of the corpus; (2) the Python interface that facilitates querying and analyzing the corpus using the NLTK and spaCy libraries and (3) a set of HTML visualizations aimed at the general public to navigate the corpus and explore differences between TF-IDF frequencies.

Keywords: Political Speech Corpus, Corpus Visualization, TF-IDF frequency

1. Introduction and Previous Work

Political speeches attract a lot of attention, both among specialists and the general public. Due to their historical and cultural significance, political speeches are subject of study in political communication, sociological studies and discourse analysis. Their availability also make political speeches a very informative piece of linguistic data to inform language use. As a result, political speech transcriptions have been widely exploited as corpus material for very diverse endeavours (Ahrens, 2006; Bevitori, 2015; Charteris-Black, 2004; Laver et al., 2003; Light, 2014; Savoy, 2010; Sim et al., 2013).

The Presidential Inaugural Addresses Corpus from the NLTK toolkit (Bird et al., 2009) (that compiles the inaugural speeches by the US presidents from 1789 to 2009) is a good example of an English corpus of political speeches that has been broadly used for educational, sociological and linguistic purposes. Efforts are being made to produce digitized and reusable corpora of political speeches in other languages other than English (Barbaresi, 2018; Osenova and Simov, 2012).

This paper introduces a new corpus of political speeches in Spanish that range from 1937 to 2019. The collected texts are all Christmas Eve National Speeches, a speech that is given every year by the king of Spain where he shortly addresses some of the public issues that affect the country. This tradition began in 1937, when fascist dictator Francisco Franco addressed the nation on New Year's Eve. The Christmas speech was celebrated yearly[1] during the following 40 years of Franco's regime and the tradition was preserved by the following heads of state after Franco's death in 1975 and continues up to this day. This collection of speeches is an interesting piece of both linguistic data and sociopolitical information, as they cover a broad period of time of Spain's recent history and portrait some of the changes on political matters, social concerns and public discourse that have taken place in Spain during the last 80 years.

Although the speech is massively followed and analyzed by the media, there is no easy or comfortable interface that allows to interact and aggregate the data from the different speeches: the speeches from 1975 onwards are published yearly by newspapers and can be found on the royal family website, but there is no available corpus that allows data retrieval and linguistic analysis of the speeches as a whole easily. The speeches from the dictatorship period (1937-1974) are scattered in fascist propaganda websites and forums and have not been digitized or made digitally available by any official institution, despite the historical interest of these documents.

In this paper we present a corpus that contains the digitized version of all the available Christmas speeches and and application to query and analyze the corpus.

This paper describes (1) the corpus of Spanish political speeches from 1937 to 2019; (2) the available Python interface to query and analyze the corpus and (3) a set HTML visualizations aimed at the general public to navigate the corpus and explore differences between lexical frequencies.

2. Corpus Description

The corpus consists of 206,937 tokens from 77 documents written in European Spanish[2]. These documents correspond to the Christmas speeches delivered yearly by the head of state between 1937 and 2019 (although there was no speech between 1939 and 1945).

The 77 speeches were retrieved through web scraping of different sources and are available as downloadable plain text files[3]. The csv file `metadata` contains the metadata associated with each speech: year in which the speech was delivered, head of the state that gave the speech, name of the file and URL from where the speech was retrieved.

[1] There was no speech between 1940 and 1945

[2] The entire corpus (as well as the Python interface and HTML visualizations described in sections 3. and 4.) have been made publicly available at `https://github.com/lirondos/discursos-de-navidad`

[3] `https://github.com/lirondos/discursos-de-navidad/tree/master/data/`

The corpus contains speeches by three different speakers, alas the three subsequent heads of state since 1937[4]: dictator Francisco Franco, king Juan Carlos I and king Felipe VI. Table 1 contains an overview of corpus split per speaker.

Years	Speeches	Tokens	Speaker
1937-1974	32	140,827	Francisco Franco
1975-2013	39	55,541	Juan Carlos de Borbón
2014-2019	6	10,569	Felipe de Borbón
Total	77	206,937	

Table 1: Number of tokens and documents per speaker

3. Corpus Interface

This section documents the programming infrastructure that facilitates querying and analyzing the corpus. This infrastructure integrates the plain text speech files described in the previous section with NLTK and spaCy libraries in order to facilitate corpus querying and linguistic data analysis.
The corpus interface provides two classes to access and query the speeches through Python: the Speech class and the Corpus class.

3.1. The Speech Class

The Speech class formalizes the structure and data related to every single speech. Every speech object has the following fields:

1. the raw text of the speech
2. the list of word tokens
3. the list of word types
4. the list of sentences
5. the list of paragraphs
6. a lemmatized and POS-tagged version of the text (provided by the Python library SpaCy)
7. a NLTK Text object version of the text
8. the year when the speech was delivered
9. the head of state it was delivered by
10. the historical period when the speech was delivered

In terms of historical periods, seven historical periods have been considered bearing in mind the historical circumstances and events that have taken place in Spain from 1937 to nowadays:

- 1937-1939: Spanish Civil War.
- 1940-1959: Early Francoism.
- 1960-1974: Late Francoism.
- 1975-1981: Transition to democracy.
- 1982-1995: Socialist period.
- 1996-2007: Economic bubble.
- 2008-2019: Economic recession.

These periods of time will become particularly relevant in the visualization process (see Section 4).
The methods within the Speech class query the speech and extract relevant lexical information from it. The Speech methods can be divided in three groups: methods that provide absolute lexical information (not filtered or ordered by frequency), methods that provide information filtered by or related to frequency and methods that query for information on a particular word in the speech.

1. Methods that return lexical information (on absolute terms):
   - length(): number of words in the speech.
   - content_words(): list of content words (words that do not appear on the stopword list).
   - bigrams(): list of bigrams.
   - trigrams(): list of trigrams.
   - content_bigrams(): list of content bigrams (bigrams where both words are content words, i.e. words not in the stopwords list).
   - content_trigrams(): list of content trigrams (trigrams where the first and third words are content words, i.e. words not in the stopwords list).
   - longest_words(): list of the longest words in the corpus.
2. Methods that return frequency distributions:
   - frequencies(): frequency distribution of all words in the speech.
   - most_frequent_content_words(): list of *(word, frequency)* pairs ordered on the frequency where *word* is a content word (words that do not appear on the stopword list).
   - most_frequent_bigrams(): list of *((word1, word2), frequency)* elements where *word1* and *word2* are content words.
   - most_frequent_trigrams(): list of *((word1, word2, word3), frequency)* elements where *word1* and *word3* are content words.
   - hapaxes(): list of hapaxes (words that only appear once in the speech).
3. Word query methods:
   - word_appearances(word): number of times that a word appears on the speech.

[4] The first two speeches in the corpus were delivered by dictator Francisco Franco in 1937 and 1938 after the *coup d'etat* that led to the Spanish Civil War. It is arguable that he can be considered the head of the state at the time, as the legitimate government of the nation remained partially in power in some areas of Spain until 1939. These speeches have been nonetheless included in the corpus due to their historic significance. Their inclusion and the denomination of the corpus as "speeches delivered by the head of the state" in no way endorse any legitimacy on the dictator.

- concordance(word): concordances of that given word (contexts in which the word appear)
- similar(word): words that tend to appear in the same contexts as the specified word.
- dispersion_plot(words): creates a dispersion plot that display the appearences of a list of words.

Methods of the Speech class
Lexical information
length()
content_words()
bigrams()
trigrams()
content_bigrams()
content_trigrams()
longest_words()
Frequency information
frequencies()
most_frequent_content_words()
most_frequent_bigrams()
most_frequent_trigrams()
hapaxes()
Word queries
word_appearances(word)
concordance(word)
similar(word)
dispersion_plot(words)

Table 2: Methods in the Corpus interface by type

3.2. The Corpus Class

The Speech object that was described in the previous section models and handles one speech only. However, the user may want to analyze several speeches at the same time: the user might want to know what are the most frequent content bigrams, not only in a given speech, but in all the speeches in a decade, or for a given head of state or during a particular time period, for instance. The Corpus class draws together several speeches, so that the same types of queries and analysis can be performed on a set of speeches (not just on a single one) and facilitates the comparison between different segments of the corpus.
The Corpus object is a collection of Speeches. The Corpus object is highly customizable: a Corpus object can contain all the speeches in the speeches folder (all speeches from 1937 to 2019), only the speeches of a given period of time, or even build an object that contains only one speech or a selected choice of speeches. The Corpus class offers different methods to call the Corpus constructor depending on which of these options is chosen.
Every Corpus object consists of:

- A PlaintextCorpusReader object from the NLTK library.
- A dictionary of Speeches, where every year in the corpus is a key and its value is the Speech object for that year. This means that every speech on the corpus has its own Speech object.
- A list of the years covered by the Corpus object.
- A Speech object that combines all the speeches in our selection into a Speech object. This means that all the speeches in our selection are concatenated and treated as a very long speech. This very long Speech object facilitates applying the same queries that we did for a Speech object on various speeches at the same time.
- A Complementary Corpus. This object is only created when the Corpus object does not contain the entire set of speeches. The Complementary Corpus is a type of Corpus that contains all the speeches that are not in the current selection of our Corpus. This field enables to know what words, bigrams, trigrams etc that appear in our Corpus are unique to that subset of the corpus, i.e. they do not appear in the Complementary Corpus.
- The unique_words field contains all the words that are in our current Corpus that are unique to the speeches in the Corpus, that is, words that are not found in the Complementary Corpus.

The most important method for corpus analysis is the radiography() method. Given a Corpus object, radiography() returns relevant lexical information about that Corpus. If the Corpus was built using all the speeches, the information returned by the radiography() method will concern the entire period of time from 1937 to 2019. If only a subset of speeches (or even just one) was used to build the Corpus object, then the returned information will concern only those speeches.
The radiography() method displays the following information:

- Years included in the corpus.
- Total number of speeches in the corpus.
- Total number of words.
- Words per speech ratio.
- Frequency distribution for the words in the corpus.
- Frequency distribution for content words.
- Hapaxes contained in the corpus.
- Words that are unique to the corpus, that is, words that appear exclusively on the created corpus, but not in the Complementary Corpus (if there is no Complementary Corpus because the created Corpus contains all the speeches a message is displayed).
- Frequency distribution for content bigrams.
- Frequency distribution for content trigrams.

4. Corpus Visualizations

Finally, we present a collection of HTML visualizations that facilitate navigating the corpus and exploring the differences in word usage between different time periods. These visualizations are aimed at non-technical users (journalists, analysts, general public, etc) that may be interested in querying the corpus but that might lack the technical knowledge to use the Corpus and Speech classes in Python.

The HTML visualizations were created using Scattertext (Kessler, 2017), a browser-based library that visualizes how two corpora differ. Scattertext splits a given corpus of texts into two subcorpora and produces interactive visualizations that display the lexical differences between the two subcorpora and allows to see word appearances in context.

The political speech corpus can be divided into several subcorpora based on different distinctive features:

- Based on the head of the state delivering the speech (dictator Francisco Franco, king Juan Carlos I, king Felipe VI).
- Based on the political system, we can split the corpus into two halves: Spain under the francoist dictatorship (1937-1977), democratic Spain (1978-2019).
- Based on the chronological time when the speech was delivered, we can split the corpus into different periods: Spanish Civil War (1937-1939), Early Francoism (1940-1959), Late Francoism (1960-1974), Transition to democracy (1975-1981), Socialist period (1982-1995), Economic Bubble (1996-2007) and Economic Recession (2008-2019).

The `visualize.py` script splits the Corpus containing the entire collection of speeches into these different groups and feeds them to the Scattertext library. The Scattertext library produces HTML interactive visualizations that portray the main lexical differences between the two subcorpora according to TF-IDF frequencies, plot them on an Cartesian axis, facilitates word search and displays word concordances.

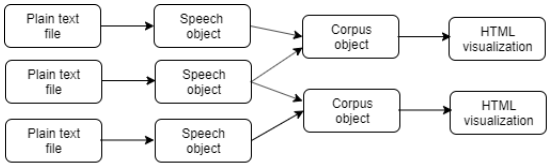

Figure 1: Diagram flow with the connections between the corpus text, the Python integration and the HTML visualizations.

For example, it is possible to provide Scattertext with two subcorpora from the speeches corpus, one subcorpus with all speeches from the fascist period (1937-1977) and the other subcorpus with all the speeches from the democratic period (1978-2019). Scattertext will measure the frequency information in terms of the TF-IDF values and display the differences in word usage on a graph. The y axis will display the TF-IDF values for one of the subcorpus, the x axis will display the TF-IDF values for the other. Consequently, words that are displayed with a high value for both x and y are words that have high TF-IDF values in both subcorpora. On the other hand, words that are displayed with a low value for x and a high value for y are words that have low TF-IDF values in one subcorpus but a high value in the second subcorpus. Words that are near the diagonal are words that have similar TF-IDF values in both subcorpora.

Eleven HTML visualizations have been produced integrating the Corpus class and Scattertext[5]. These visualizations have been produced according to the possible subdivisions listed above: head of state (dictator Francisco Franco, king Juan Carlos I, king Felipe VI), historical periods (Spanish Civil War, Early Francoism, Late Francoism, Transition to Democracy, Socialist Government, Economic Bubble, Economic Recession) and government system (speeches under fascist regime, prior to 1978; speeches during democracy, 1978 onwards). The HTML visualizations also include the possibility of localizing a certain word on the graph and displaying its concordances and frequency values (note that not all words on the corpus make it to the visualization, only the most significant words are displayed). These visualizations can help contextualize how the Spanish political discourse has changed during the last decades, and the lexical changes that have taken place in the speeches of the heads of state during the last 80 years.

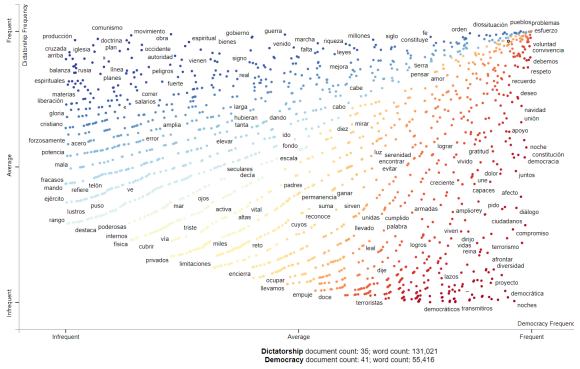

Figure 2: TF-IDF frequency comparison between speeches before and after 1978.

5. Conclusions and Future Work

We have presented a new corpus of political speeches in Spanish language. This corpus consists of the Christmas speeches to the nation by the Spanish head of state since 1937. The corpus includes the digitized texts of 77 speeches, along with metadata information about the year, speaker and the data source where the text was retrieved. To our very best knowledge, this is the first time these speeches have been compiled into a single resource in a digital and reusable fashion.

Secondly, we have introduced a Python interface that integrates the texts with NLTK and spaCy utilities and that allows corpus exploration and comparison of frequencies and collocations along different time periods.

[5] Live visualizations available at `https://lirondos.github.io/discursos-de-navidad/`

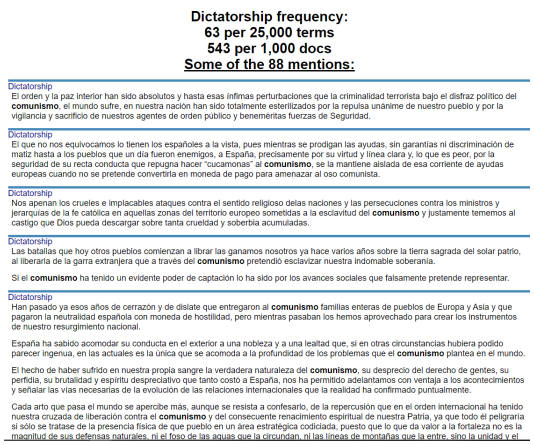

Figure 3: Concordances visualization of the word *comunismo* ("comunism").

Lastly, we have also produced several interactive HTML visualizations of lexical frequencies of the corpus based on TF-IDF measures. These visualization have been developed using Scattertext library and facilitate corpus navigation and comparison for non-technical users.

In terms of future work, this project is an ongoing project: future speeches will continue to be added to the corpus as long as the traditional Christmas speech lasts.

6. Bibliographical References

Ahrens, K. (2006). Using a small corpus to test linguistic hypotheses: Evaluating "people" in the state of the union addresses. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 11, Number 4, December 2006, pages 377–392.

Barbaresi, A. (2018). A corpus of German political speeches from the 21st century. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 792–797, Paris, France. European Language Resources Association (ELRA).

Bevitori, C. (2015). Discursive constructions of the environment in american presidential speeches 1960–2013: A diachronic corpus-assisted study. In *Corpora and Discourse Studies*. Springer, pp. 110–133.

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

Charteris-Black, J. (2004). Why "an angel rides in the whirlwind and directs the storm": a corpus-based comparative study of metaphor in British and American political discourse. In *Advances in Corpus Linguistics*. Brill Rodopi, pp. 133–150.

Kessler, J. (2017). Scattertext: a browser-based tool for visualizing how corpora differ. In Proceedings of ACL 2017, System Demonstrations, pages 85–90, Vancouver, Canada, July. Association for Computational Linguistics.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American political science review*, 97(2):311–331.

Light, R. (2014). From words to networks and back: Digital text, computational social science, and the case of presidential inaugural addresses. *Social Currents*, 1(2):111–129.

Osenova, P. and Simov, K. I. (2012). The political speech corpus of bulgarian. In LREC, volume 2012, pages 1744–1747. Citeseer.

Savoy, J. (2010). Lexical analysis of us political speeches. *Journal of Quantitative Linguistics*, 17(2):123–141.

Sim, Y., Acree, B. D., Gross, J. H., and Smith, N. A. (2013). Measuring ideological proportions in political speeches. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 91–101.