

The CompWHoB Corpus: Computational Construction, Annotation and Linguistic Analysis of the White House Press Briefings Corpus

Fabrizio Esposito[†], Pierpaolo Basile[‡], Francesco Cutugno^{*}, Marco Venuti^{*}

[†] Dept. of Humanities, University of Naples Federico II

[‡] Dept. of Computer Science, University of Bari Aldo Moro,

^{*} DIETI, University of Naples Federico II,

^{*} Dept. of Humanities, University of Catania

fabrizio.esposito3@unina.it, pierpaolo.basile@uniba.it,
francesco.cutugno@unina.it, mvenuti@unict.it

Abstract

English. The CompWHoB (Computational White House press Briefings) Corpus, currently being developed at the University of Naples Federico II, is a corpus of spoken American English focusing on political and media communication. It represents a large collection of the White House Press Briefings, namely, the daily meetings held by the White House Press Secretary and the news media. At the time of writing, the corpus amounts to more than 20 million words, covers a period of time of twenty-one years spanning from 1993 to 2014 and it is planned to be extended to the end of the second term of President Barack Obama. The aim of the present article is to describe the composition of the corpus and the techniques used to extract, process and annotate it. Moreover, attention is paid to the use of the Temporal Random Indexing (*TRI*) on the corpus as a tool for linguistic analysis.

Italiano. *Il CompWHoB Corpus, in sviluppo presso l'Università di Napoli Federico II, è un corpus di parlato inglese-americano comprendente le conferenze condotte dai segretari statunitensi per i rapporti con la stampa, definite come Press Briefings. Allo stato attuale il corpus è composto da più di 20 milioni di parole e si estende dal 1993 sino a fine 2014. L'obiettivo di questo articolo è di descrivere la composizione del corpus, le tecniche utilizzate per estrarre ed annotare i testi, e mostrare come possa fungere da fonte di analisi linguistica attraverso l'utilizzo del Temporal Random Indexing (*TRI*).*

1 Introduction

As political speech has been gaining more and more attention over recent years in the analysis of communication strategies, political corpora have become of paramount importance for the fulfilment of this objective. The CompWHoB Corpus, a spoken American English corpus currently being developed at the University of Naples Federico II, wants to meet the need for political language data, as it focuses on the political and media communication genre. This resource is a large collection of the transcripts of the White House Press Briefings, namely, the daily meetings held by the White House Press Secretary and the news media. As one of the main official channels of communication for the White House, briefings play indeed a crucial role in the administration communication strategies (Kumar, 2007). The corpus currently amounts to more than 20 million words and spans from 1993 to 2014, thus covering a period of time of twenty-one years and five presidencies. Work is underway to extend the corpus so as to reach the end of the second term of President Barack Obama. Unlike other political corpora such as CORPS (Guerini et al., 2008; Guerini et al., 2013) and the Political Speech Corpus of Bulgarian (Osenova and Simov, 2012), the CompWHoB does not include monological situations, due to the inherent dialogical characteristics of the briefings. As other web corpora (Baroni and Kilgarriff, 2006; Baroni et al., 2009; Lyding et al., 2014), the CompWHoB can be considered a web corpus (Kilgarriff and Grefenstette, 2003; Hundt et al., 2007), since its texts are directly extracted from *The American Presidency Project* website. Moreover, it should be pointed out that WHoB is a pre-existing specialized corpus (Spinzi and Venuti, 2013) annotated by using XML mark-up and mainly employed in the field of corpus lin-

CompWHoB Corpus							
Presidency	texts	tokens	tokens mean	types	TTR	turn-takings	WHo-s
Bill Clinton_1	1,072	4,581,665	4,274	79,129	36.97	116,437	497
Bill Clinton_2	1,066	4,658,054	4,370	81,789	37.89	102,160	525
George W. Bush_1	777	3,660,600	4,711	65,635	34.30	78,992	133
George W. Bush_2	1,057	4,536,616	4,292	73,809	34.65	82,702	286
Barack Obama_1	804	4,470,070	5,560	76,604	36.23	87,432	299
Barack Obama_2	463	3,344,567	7,224	48,493	26.51	44,982	74
TOTAL	5,239	25,251,572		426,458		512,651	1,814

Table 1: Composition of the CompWHoB Corpus in its current stage (July 2015); _1 and _2 stands for the first term and second term of each presidency, respectively; type-token ratio was calculated using Guiraud’s (Guiraud, 1954) index of lexical richness; WHo-s stands for White House staff, namely, personnel identified as belonging or related to the White House presidential staff.

guistics. Thus, the aim of the present article is to describe how the corpus can be used as a future resource in different research fields such as computational linguistics, (political) linguistics, political science, etc.

The paper is structured as follows: Section 2 gives an overview of the corpus. Section 3 describes the details of the corpus construction and annotation. The use of TRI on the corpus is then discussed in Section 4. Lastly, Section 5 concludes the paper.

2 Corpus Overview

The CompWHoB Corpus consists of the transcripts of the press conferences held by the White House Press Secretaries and/or other administration officials and the news media. The texts that form the corpus were all extracted from the American Presidency Project website www.presidency.ucsb.edu, where the Press Briefings document archive section can be freely consulted. Data was collected and formatted into a standardized XML encoding, according to the TEI Guidelines (Sperberg-McQueen and Burnard, 2007). In some cases, texts were subsequently split to mark the beginning of the new president first term. Six are the presidencies represented in the CompWHoB Corpus: both Bill Clinton and George W. Bush eight-year term are included, while the second term of the incumbent US President, Barack Obama, is not complete since he is currently in office. Thus, at the current stage (July 2015) the corpus contains a total of 5,239 texts comprising 25,251,572 tokens and 422,891 types, and spans from January 27, 1993 until December 18, 2014. Given the inherent dialogical characteristics of press conferences, a total

number of 512,651 turn-takings has been calculated so far. Across the time span covered by the corpus, 1,814 are the speakers individually identified as press secretaries, presidential staff members or administration officials. See Table 1 for more details.

3 Corpus Construction and Annotation

3.1 Construction and Structural Annotation

Data extracted comes in a standardized format. Each briefing consists of a transcript where every turn-taking is signalled by the use of the capital letters to identify the speaker. Two are the main roles found in the transcriptions: the podium, namely, the White House Press Secretary or any other administration official, always identified by their surnames; the press corps, identified by the use of the capital letter Q. Information about the date of the event was extracted and then added to the beginning of every press conference. As first step after data extraction, the resulting texts were encoded in XML format in a semi-automatic way by using regular expressions and manual checking. Transcripts were then mapped to XML files according to a calendar year division. Meta-textual information contained in the data was encoded as well so as to enrich the corpus and make it easily navigable. Thus, the CompWHoB Corpus is structured as follows: every year forming part of the corpus is diachronically structured. A *div* tag was created to mark the beginning and the end of every transcript. An attribute value shows the date of that specific event in a *yyyy-mm-dd* format. Every *div* contains the dialogical situation of the press conference, where each speaker is identified by the use of a *u* tag. In order to provide

an in-depth description of the sociolinguistic characteristics of the speakers, every *u* tag consists of self-explanatory multiple attributes: *role*, *sex* and *who*. Since in the transcripts press corps are only identified by the capital letter Q, it was impossible neither to recover information about the gender nor the name. Thus, for every media member the attribute value *sex* is always *u*, namely, unknown, and both *role* and *who* attribute values are always *journalist*. Conversely, since information about Press Secretaries and members related to the presidential administration staff was available in the transcripts, attribute values contain information about the role, gender and name of the speaker. This operation had to be made manually, but one of the main objectives of this work is to make it semi-automatic querying an existent political database that will make the process less burdensome. As many are the White House members involved in the press conferences, we decided to categorize them by role. Thus, Press Secretaries are the only ones identified as *podium*, due to their function of conducting the briefing. Administration officials and presidential staff members can be instead recognized by the role value *podium_* plus the position held by them (e.g. military, administration, etc.). The beginning and the end of every speech is marked by the use of *p* tags. As original transcripts contained also meta-textual information enclosed in brackets about audience reactions and speech events descriptions (e.g. (Laughter), (Applause), etc.), we decided to keep it so as to broaden and vary future analysis approaches. See Table 2 for a summary of these tags. See Table 3 for the description of the corpus press conference structure.

Tag
{event type="laughter"}
{event desc="applause"}
{event desc="inaudible"}

Table 2: Meta-textual speech events tags

{div1} # date of the press conference
{u} # identification of the speaker
{p} # speech of the identified speaker
{self-closing tag} # extra-textual speech events

Table 3: CompWHoB briefing structure

3.2 Linguistic Annotation

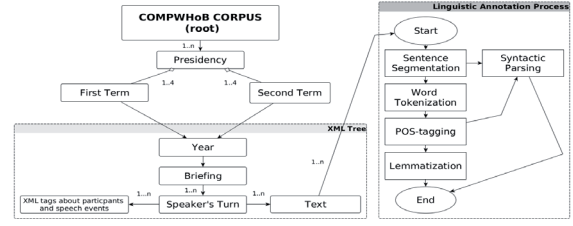


Figure 1: CompWHoB structure and linguistic annotation process

As regards the NLP aspect (Figure 1), we chose to adopt Python (3.4 version) as programming language, using the Natural Language Toolkit (NLTK) platform (Bird et al., 2009), since it provides a large suite of libraries for natural language processing. As first step, sentence segmentation and word tokenization were carried out. POS-tagging was then performed employing the Penn Treebank tag set (Marcus et al., 1993), trained on the Treebank Corpus. We made this choice to have immediately a first grasp on the linguistic data. Being at the early stages of our work, we decided to test NLTK POS tagger by comparing the output with a human-labeled Gold Standard test set consisting of 24 sections randomly selected from the corpus, amounting to over 500 tokens. Since at the current stage POS tagging achieves an accuracy of 92%, our future aim is to improve the performance of NLTK POS tagger once the corpus is complete, providing it with a syntactic parsing as well. As for the lemmatization of the resulting texts, we decided to use the WordNet lemmatizer provided by the NLTK platform. During this task we had to map the part-of-speech tags to the WordNet part-of-speech names in order to get a more accurate output. Texts processing tasks were always performed taking into account each turn-taking. This means that, at the current status, one of the main advantages of the CompWHoB Corpus is the possibility to retrieve linguistic information by specifying the name and/or the role of the speaker, allowing an in-depth analysis of the acquired information. This is why our primary objective in the near future is to provide the means to query the corpus. We plan to reach this goal by employing the Corpus Workbench (CWB) architecture and the *Corpus Query Processor* (Christ et al., 1999; Evert and Hardie, 2011).

4 TRI on the CompWHoB Corpus

Our intention was to perform a linguistic analysis with the aim of finding some variation in word usage across several presidential and political mandates. We chose to model word usage exploiting distributional semantic models (Sahlgren, 2006). In a distributional semantic model, words are represented as mathematical points in a geometric space. Similar words are represented close in that space. The space is built taking into account words co-occurrences in a large corpus. One drawback of this kind of approach is that geometric spaces built on different corpora are not comparable. Moreover the temporal feature is not included in these models. Considering the peculiarities of the CompWHoB Corpus such as temporal information and different speakers, a technique able to manage these kind of features is needed. Recently, a technique called TRI based on Random Indexing (Sahlgren, 2005) able to manage temporal information has been proposed in (Basile et al., 2014). TRI can build different word spaces for several time periods allowing the analysis of how words change their meaning over time. Relying on TRI, we build six separate word spaces, one space for each presidency. The first goal of our analysis is to find interesting words that change their meaning across time. Since word vectors in each word space are made comparable thanks to the TRI tool, it is possible to compare the similarity of a word vector in each word space. In particular, given a word w and two time periods t_1 and t_2 is possible to compare the cosine similarity between the word vector of w in t_1 and word vector of w in t_2 . A low level of similarity between vectors indicates a high word usage variation across the two time periods. Exploiting this technique we discovered some words that significantly change their usage. In this case, it is worth paying attention to the words resulting from the time periods representing the end of a presidency second term and the beginning of a new one. For example, investigating the neighbourhood of the word *Guatemala* in *Clinton_2/Bush_1*, we note that in *Clinton_2* words such as *donors*, *accord* and *workable* appear, while in *Bush_1* the word *Guatemala* is near to other geo-political entities, for example: *honduras* and *slovak*. Investigating historical events in that period we found that in 1999 President Clinton finally apologized for America's role in almost a half-century of re-

pression in Guatemala.

The second analysis concerns how a particular topic is treated. We selected the topic of the American debate on guns. The idea was to analyse how each presidency discusses this subject. We selected the word *gun* as the representative word of the topic. Moreover, we expanded the topic employing semantic frames in which the word *gun* had been previously used. We adopted FrameNet to extract relevant frames. Following this methodology we identified other relevant words: *firearm*, *handgun*, *machine-gun*, *shooter*, *shotgun* as nouns; and *discharge*, *fire*, *hit*, *shoot* as verbs.

In order to represent the gun topic in the word space we adopted the vector sum operator. For each word space a vector was built, representing the vector sum of words belonging to the topic. The sum vector is used to retrieve the most similar vectors using cosine similarity. This operation was repeated for each administration. The idea was to analyse the neighbourhood of the gun topic in each presidency. Results show a clear evolution in how the different administrations dealt with this subject. While in Bill Clinton and George W. Bush presidencies the first fifteen most similar vectors mainly denote the semantic field of weapons, it is only from the Obama administration that adjectives and nouns appealing to emotions make their appearance (e.g. *heartening*, *suffer*, *grassroots*, *darn*), marking a new era in the White House communication strategies about the gun issue.

5 Conclusions

At the time of writing, the CompWHoB Corpus is probably one of the largest political corpora mainly based on spontaneous spoken language. This feature represents one of its strongest points, as the linguistic analysis performed by employing the TRI has proved. As for the near future, two are our main goals: the first one is to make the process of structural annotation as much computational as possible by retrieving information from available political databases; the second one is to provide the corpus with syntactic parsing and improve the overall performance of the linguistic annotation process. In terms of accessibility, we intend to make the CompWHoB Corpus available via the CPQ web interface (Hardie, 2012) by the end of next year. For now, the fully annotated corpus is accessible and available on request.

References

- Marco Baroni and Adam Kilgariff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 87–90. East Stroudsburg.
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226, September.
- Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing Word Meaning over Time by Exploiting Temporal Random Indexing. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *First Italian Conference on Computational Linguistics CLiC-it 2014*. Pisa University Press.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.
- Oliver Christ, Bruno M. Schulze, and Esther Knig. 1999. *Corpus Query Processor (CQP). User's Manual*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, Germany.
- Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, Birmingham, UK. University of Birmingham.
- Marco Guerini, Carlo Strapparava, and Oliviero Stock. 2008. Corps: A corpus of tagged political speeches for persuasive communication processing. 5(1):19–32.
- Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. *The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions*, volume 7688 of *Lecture Notes in Computer Science*, pages 86–98. Springer Berlin Heidelberg.
- Paul Guiraud. 1954. *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie*. Presses Universitaires de France, Paris.
- Andrew Hardie. 2012. Cqpweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17:380–409.
- Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer. 2007. *Corpus linguistics and the web*. Rodopi, Amsterdam and New York.
- Adam Kilgariff and Gregory Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–347.
- Martha J. Kumar. 2007. *Managing the Presidents Message: the White House Communications Operation*. The John Hopkins University Press.
- Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISÀ Corpus of Italian Web Texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden, April. Association for Computational Linguistics.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.
- Petya Osenova and Kiril Simov. 2012. The political speech corpus of bulgarian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Magnus Sahlgren. 2005. An Introduction to Random Indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5.
- Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm: Stockholm University, Faculty of Humanities, Department of Linguistics.
- C. Michael Sperberg-McQueen and Lou Burnard. 2007. TEI P5: Guidelines for Electronic Text Encoding and Interchange.
- Cinzia Spinzi and Marco Venuti. 2013. *Tracking the change in an Institutional Genre: A Diachronic Corpus-based study of White House Press Briefings*, pages 182–197. Cambridge Scholars Publishing, Newcastle upon Tyne.