A Multilingual Corpus of German, French and Italian Political Discourse: Goals and Methodological Challenges

Silvia Modena, Marcella Palladino, Vincenzo Gannuscio

Università degli Studi di Modena e Reggio Emilia
{silvia.modena \ marcella.palladino \ vincenzo.gannuscio}@unimore.it

Abstract: This contribution offers a first overview of a work-in-progress project focused on the creation of a multilingual corpus of political discourse. The languages involved in the project are German, French and Italian, and the data contained in the corpus are both (digital) written and spoken texts. The corpus is meant to be a collaborative source for the politolinguistic scientific community in order to facilitate the process of data collection. The selection of the texts is founded on the definition of 'populism', since the further aim of the project is to investigate the rise of populist discourse from 2022 until the present days. However, the project can be expanded in the future to involve the entire political spectrum and all types of political actors. The planned methods for the creation of the corpus as well as their potential limitations are delved into and the expected results are presented.

1 Introduction

The multilingual corpus presented in this contribution is a collection of political texts under construction and undergoing expansion, grounded on a project of the University of Modena and Reggio Emilia called PO.POL.I (POpulismo - POLitica - Identità). Its aim is to diachronically investigate the rise of populist discourse in Germany, France and Italy from 2022 to the present. To achieve this goal, a corpus of political (digital) written texts (among others, hashtag, link, @) and orthographic transcripts of German, French and Italian non-institutional political spoken texts will be built.

The project is based on a theoretical background that combines several linguistic approaches: Frameworks of analysis from the German [1] [2] and Italian [3] politolinguistics as well French studies on discourse analysis [4] [5] are acknowledged. This project aims to offer a collaborative collection of data that is available and can be implemented by the users of the corpus. This data collection, made available in several output formats, serves as a basis to carry out analyses on the populist use of political language such as for instance prosodic, metaphorical, lexical and lexicometrical analysis.

The paper will delve into the creation of the corpus together with the planned methods used to build the corpus and to conduct the analysis planned for the project. The expected results will be discussed and the limitations will be mentioned. The corpus is in its initial stages and future studies are foreseen to further define and expand the presented methods and goals. Due to the collaborative nature of this project, which addresses mainly the politolinguistic scientific community, methodological contributions and discussions will still remain open.

2 Corpus and methodology

The corpus to be compiled consists of written and spoken data, and its creation involves the use of various tools and software programs. The parameters to select the materials for the corpus revolve around the populist character of the data. 'Populism' is considered as a broad concept, not necessarily linked to certain parties or politicians, but rather as a linguistic and communicative tendency of (political) written and spoken texts. The term 'populism' refers in this contribution to the definition given by Mudde & Kaltwasser [6], namely "*a thin-centered ideology that considers society to be ultimately separated into two homogeneous and antagonistic camps, "the pure people versus "the corrupt elite", and which argues that politics*

should be an expression of the volonté générale (general will) of the people"[1]. Populism, according to this definition, is an ideology that can be identified in several written and spoken texts, without being limited to specific parties. This definition can be related to Teguieff's [7] definition of "populisme protestataire-sociétal" ('protest and societal populism'), which also considers populism as an ideology. All the texts selected for the corpus share the populist character in the sense that they do contain elements that (explicitly or implicitly) support aspects of the mentioned ideology.

At the moment, the corpus consists approximately of 600.000 tokens, but the final size of the corpus still needs to be defined, and the materials will in any case be implemented by further users. This implies that the corpus is meant to be an open work, which will be expanded throughout the years. The steps for the building of the corpus are visually summarized in Fig. 1. However, the illustrated steps are still under discussion and improvement at this stage of the project, and feedback will be asked in the following months to potential users.

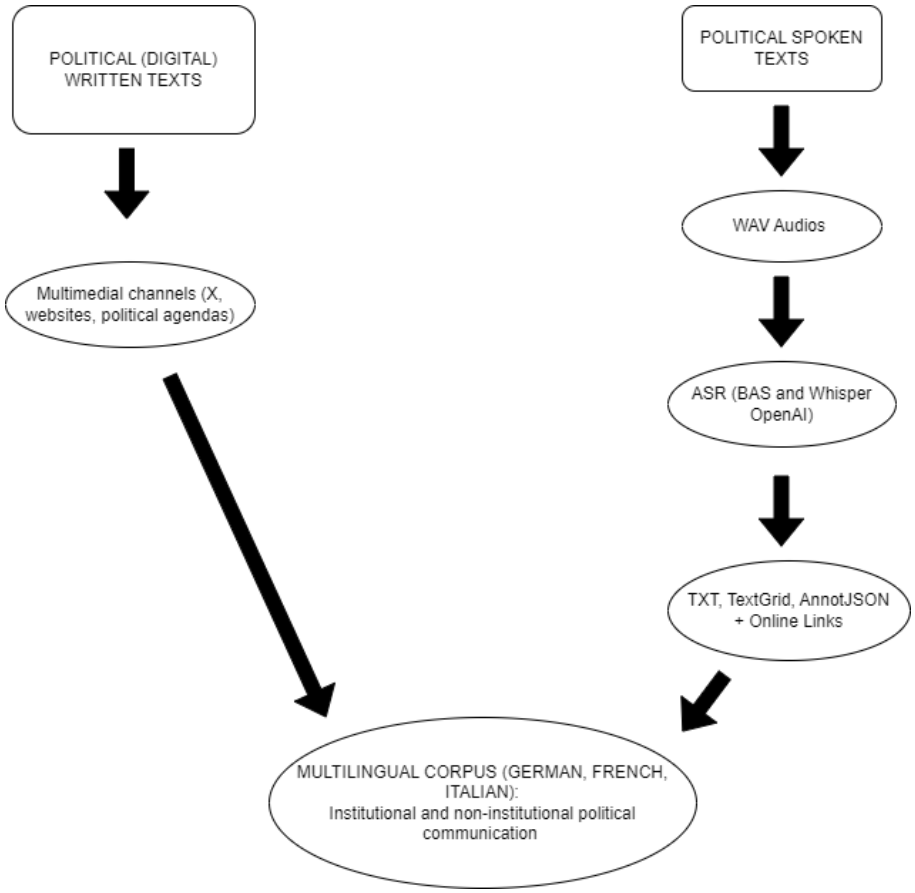

Figure 1 - Planned steps for the creation of the corpus

2.1 (Digital) written texts

Written texts fulfill three essential criteria applied in the chosen methodological framework. First and foremost, they embody a key aspect of political rituals rooted in the movements and activities of candidates during their presidential campaigns. Secondly, these texts are situated within a defined temporal framework that aligns with the electoral contest (from January 2022), a period that generates a copious discursive activity. This activity aligns with the concept of the *moment discursif* ('discursive moment'), as outlined by Moirand [8]. While this notion retains a direct connection to the broader concept of *hétérogénéité énonciative* ('enunciative

[1] Italics in the original text.

heterogeneity'), a third criterion has been opted for: The predominantly monological nature of the texts delivered by each candidate.

Initially, the written texts selected for the project excluded other forms of discursive production found in broader corpora, such as tweets (now X posts) and Facebook posts. However, recognizing the "impossible closure of media corpora" [9] from both enunciative and interdiscursive perspectives, the project will include these forms of communication. This broader approach contributes to provide a more comprehensive and systematic examination of the linguistic and rhetorical features typical of political discourse, taking into account the interplay between traditional and digital media discourses.

2.2 Spoken texts

The spoken texts selected for the corpus belong to non-institutional political discourse, namely spoken texts delivered outside the parliament. The reasons for this choice are twofold. First of all, the non-institutional speeches and interviews are generally not officially transcribed and their orthographic transcripts are not available for the scientific community. Secondly, speeches such as party rallies or election speeches are more likely to show a less institutionalized structure of communication, since their objective is generally to gain voters or consent by a more varied audience than in the parliament. For this reason, it can be assumed that populist frames of communication may be more numerous, or at least more easily identifiable in this type of spoken texts.

Political speeches and interactions will be transcribed with systems of Automatic Speech Recognition (ASR). The tools of the BAS (Bavarian Archive for Speech Signals) [10] and Whisper OpenAI [11] play the biggest role in the transcription. BAS Web Services were already applied to politolinguistics for German and Italian speeches [12], but not yet for French language; moreover, debates, interviews and speeches involving a continuous interaction of the speakers were taken into account only to a limited extent. The performance of the web tools in the orthographic transcription of political speeches was tested in a pilot study by Draxler [13].

The aim of the present project is not only to create the corpus, but also to test the methods of its creation and thus to be able to design a consistent procedure for the gathering and analysis of written and spoken data. More specifically, the outputs that are contained in the corpus are TXT, TextGrid, annot.json. The online links of the spoken sources will also be provided in the corpus.

3 Expected results and limitations

The expected results of the project will mainly depend on the further stages of the project: On the one hand, the project is meant to create a corpus that is available for politolinguistic research on (digital) written and spoken texts. On the other hand, the performance of the different tools and software programs will be tested in order to determine which ones appear to be the more suitable for the creation of a collection of political texts. Moreover, the tests will also concern the automatization of orthographic transcription and the different ASR systems that are taken into account.

Consistently with the aim of building an available corpus of political texts, the project will also shed light on some potential analyses that can be carried out using the corpus. Illustrating potential applications of the data in actual analyses can thus help the target users to observe the implications of the collection of data that can be employed in their studies.

Due to the preliminary nature of this project, several aspects have to be further defined. The planned methodology can be improved with suggestions and discussions with the scientific community in order to adapt it to the research needs of the potential users of the corpus. Moreover, technical requirements related to the ASR systems as well as the output formats of the data will be further delved into and specified in the future. In addition, which data can be made openly available and on which platform are issues under discussion.

4 Conclusion and further studies

This contribution has focused on the presentation of an ongoing project aimed at creating a multilingual corpus of political discourse. As already mentioned, both (digital) written and spoken texts are considered, and the corpus is meant to be further expanded by the users as well. The languages considered are German, French and Italian, but future projects involving other languages are planned. Further research will also go beyond the specific focus on populism to encompass a broader spectrum of political communication that does not necessarily exhibit the characteristics of this ideological and communicative stance. This expanded scope will allow a more comprehensive analysis of political discourse, examining how different communicative strategies, rhetorical devices, and narrative frameworks are employed across various political contexts and ideologies.

The methodology will be improved to meet the research necessities of the scientific community and the collaborative nature of the project assures that the corpus will be steadily expanded also in the following years, gathering updated texts. The potential of this project lies mainly in the possibility for the politolinguistic scientific community to have an available collection of data that they can use for their analyses. In this way, they may avoid time-consuming steps such as the orthographic transcription of spoken texts and use data that have already been checked and formatted for their scientific scopes.

Acknowledgments

Parts of the presented work have been founded by the University of Modena and Reggio Emilia, under grant FAR 2024 – Fondo di Ateneo per la Ricerca 2024: “Azione 4: Corpora multilingue e innovazione tecnologica nell’analisi del discorso populista orale”, CUP E93C24000500005.

References

[1] NIEHR, T.: *Einführung in die Politolinguistik. Gegenstände und Methoden*. Vandenhoeck & Ruprecht, Göttingen, 2014.

[2] GANNUSCIO, V.: *Wir sind das (echte) Volk. Sprachliche Ausgrenzungsstrategien der rechtspopulistischen Propaganda der AfD und der Lega Nord*. In: SCHIEWE, J., T. NIEHR, and S. M. MORALDO (Eds.): *Sprach(kritik)kompetenz als Mittel demokratischer Willensbildung. Sprachliche In- und Exklusionsstrategien als gesellschaftliche Herausforderung*, P. 43 – 61. Hempen Verlag, Bremen, 2019.

[3] CEDRONI, L.: *Politolinguistica. L’analisi del discorso politico*. Carocci, Roma, 2014.

[4] TAGUIEFF, P. A.: *L’illusion populiste: Essai sur les démagogies de l’âge démocratique*. Flammarion, Paris, 2007.

[5] MODENA, S., and L. SINI: *Les métaphores racistes dans les discours d’extrême-droite en France et en Italie - Les cas de C. Taubira et de C. Kyenge*, P. 1 – 16. Publifarum, 2015.

[6] MUDDE, C., and C. R. KALTWASSER: *Populism. A very short introduction*. Oxford University Press, 2017.

[7] TAGUIEFF, P. A.: *Le Nouveau National-Populisme*. Paris, Éditions du CNRS, 2012.

[8] MOIRAND, S.: *Les discours de la presse quotidienne. Observer, analyser, comprendre*. PUF (Linguistique nouvelle), Paris, 2007.

[9] MOIRAND, S.: *L'impossible clôture des corpus médiatiques: la mise au jour des observables entre catégorisation et contextualization*. In: *Travaux neuchâtelois De Linguistique*, 40, P. 71 – 92. 2004.

[10] KISLER, T., U. D. REICHEL, and F. SCHIEL: *Multilingual processing of speech via web services*. In *Computer Speech & Language*, Vol. 45, P. 326 – 347. 2017.

[11] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. In *International Conference on Machine Learning*, P. 28492 – 28518. 2023.

[12] PALLADINO, M.: *Webbasierte Tools für die Transkription und Analyse von Reden. Hilfreiche Instrumentarien für die (Polito)Linguistik*. In *Lingue e Linguaggi*. Università del Salento, P. 413 – 437. 2024.

[13] DRAXLER, C.: *Analysis of transcriptions using Octra – a pilot study*. In: C. DRAXLER (Ed.): *Elektronische Sprachsignalverarbeitung 2023*, P. 17 – 23. TUDpress, Dresden, 2023.