Preferred category: TALK

Methodological challenges in the creation of a corpus of German and Italian non-parliamentary political spoken communication: Systems of Automatic Speech Recognition (ASR) and linguistic research questions

Political language; non-parliamentary spoken communication; Automatic Speech Recognition; politolinguistics; German-Italian corpus

This contribution aims to illustrate challenging elements in the creation of a corpus of political spoken communication that takes place outside of the parliament and which thus embodies a less institutionalized form of communication. In disciplines such as politolinguistics (Burkhardt, 1996; Cedroni, 2014; Niehr, 2014) or political discourse analysis (Van Dijk, 1997; Spieß, 2020), not many studies have dealt with this type of communication and studies on the application of ASR-systems and its limitations for this communication seem to be only preliminary and related to speeches delivered by a main speaker (Draxler, 2023; Palladino, 2024).

Parliamentary communication is generally well-structured and speeches are often transcribed in advance, or corrected afterwards by stenographers (see, for instance, Brambilla, 2007). On the contrary, less institutionalized forms of communication may preserve a more spontaneous style and are often focused on persuading an audience of common people (to gain or maintain consent). Furthermore, recordings of non-parliamentary communication are generally publicly available on platforms such as YouTube, but the audio's quality as well as possible interactions make this type of communication not so easy to be orthographically transcribed. In addition, non-parliamentary speeches delivered for election campaigns or party rallies may also last hours and this poses a further challenge in the selection of the most adequate ASR-system. All these elements represent difficulties for researchers in the field of spoken political language who may be interested in investigating a less institutionalized spoken communication and need transcripts to conduct their analyses.

The focus of this project is on the different systems that can be useful to create a multilingual corpus of non-parliamentary political spoken communication, taking Italian and German languages under a contrastive perspective. Aspects such as interjections, hesitations, pauses, speakers' "mistakes" and false starts are considered in order to determine which challenges the ASR may solve and which further hurdles it may create. Studies on the performance of ASR for the creation of corpora were already carried out for instance by Gorisch/Gref/Schmidt (2020) and Gorisch/Schmidt (2024), showing possible limitations of ASR. However, they dealt with corpora of a different genre of speeches and especially with conversations.

Examples of different methods of ASR and their outputs will be presented and the comparison between Italian and German will be shown. These two languages were selected as a first comparison, but the future aim is to involve also further languages. The illustration of the samples from the corpus will revolve around the research necessities of linguistic analyses, which means that more technical parameters such as word error rate (WER) will not be deepened. Instead, recurrent units of analysis from politolinguistics (see, among others, Dieckmann, 2005; Girnth, 2015) and discourse analysis (see, among others, Fairclough/Fairclough, 2012) are the focus of the discussion. The project is in its preliminary phases and the discussion with researchers from correlated disciplines will improve the methodology and give the impulse to consider further relevant units of analysis.

Cited References:

Brambilla, Marina M. (2007): *Il discorso politico nei paesi di lingua tedesca: metodi e modelli di analisi linguistica*. Roma: Aracne.

Burkhardt, Armin (1996): Politolinguistik: Versuch einer Ortsbestimmung. In: Klein, J., Diekmannshenke, H. (Ed.): *Sprachstrategien und Dialogblockaden: Linguistische und politikwissenschaftliche Studien zur politischen Kommunikation*. Berlin/Boston: De Gruyter. P. 75-100.

Cedroni, Lorella (2014): *Politolinguistica. L'analisi del discorso politico*. Roma: Carocci Editore.

Dieckmann, Walther (2005): Demokratische Sprache im Spiegel ideologischer Sprach(gebrauchs)konzepte. In: Kilian, J. (Ed.): *Sprache und Politik. Deutsch im demokratischen Staat*, Vol. 6. Mannheim: Dudenverlag. P. 11–30.

Draxler, Christoph (2023): Analysis of transcriptions using Octra – a pilot study. In: Draxler, C. (Ed.): *Elektronische Sprachsignalverarbeitung 2023*, 105. Dresden: TUDpress. P. 17–23.

Fairclough, Isabela, Fairclough Norman (2012): *Political Discourse Analysis: A Method for Advanced Students*. London/New York: Routledge.

Girnth, Heiko (2015): *Sprache und Sprachverwendung in der Politik. Eine Einführung in die linguistische Analyse öffentlich-politischer Kommunikation*. Berlin/Boston: De Gruyter.

Gorisch, Jan, Gref, Michael, Schmidt, Thomas (2020): Using Automatic Speech Recognition in Spoken Corpus Curation. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France. European Language Resources Association. P. 6423–6428.

Gorisch, Jan, Schmidt, Thomas (2024): Evaluating Workflows for Creating Orthographic Transcripts for Oral Corpora by Transcribing from Scratch or Correcting ASR-Output. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia. ELRA and ICCL. P. 6564–6574.

Niehr, Thomas (2014): *Einführung in die Politolinguistik. Gegenstände und Methoden*. Göttingen: Vandenhoeck & Ruprecht.

Palladino, Marcella (2024): Webbasierte Tools für die Transkription und Analyse von Reden. Hilfreiche Instrumentarien für die (Polito)Linguistik. In: *Lingue e Linguaggi*, 65 (2024). Università del Salento. P. 413-437.

Spieß, Constanze (2020): Politiksprache und politische Kommunikation. In: Niehr, T., Kilian, J., Schiewe, J. (Ed.): *Handbuch Sprachkritik*. Stuttgart: Metzler. P. 302–309.

Van Dijk, Teun A. (1997): What is political discourse analysis? Universiteit von Amsterdam.