AUßERPARLAMENTARISCHE POLITISCHE KOMMUNIKATION: DATENERHEBUNG UND ANALYSEPERSPEKTIVEN

Marcella Palladino, Vincenzo Gannuscio

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

Kurzfassung[1]: Der vorliegende Beitrag stellt ein Korpus außerparlamentarischer politischer Kommunikation auf Deutsch, Französisch, Italienisch und Spanisch vor. Die Daten stammen aus der gesprochenen Sprache und wurden aus öffentlichen freien Online-Quellen erhoben. Im Fokus stehen die Datenerhebung und die möglichen Analyseperspektiven, die sich mittels des Korpus eröffnen. Transkripte außerparlamentarischer gesprochener Kommunikation sind i. d. R. schwer zugänglich und bedürfen einer relevanten Bearbeitung, damit ein homogenes Korpus erstellt werden kann. Zu diesem Zweck haben wir ein methodisches Verfahren ausgewählt, welches erlaubt, multilinguale mündliche Dateien zu transkribieren bzw. zu analysieren. Der kollaborative Charakter des Korpus sowie seine potenziellen Anwendungen werden im Beitrag ebenfalls hervorgehoben. Erste Ergebnisse liegen bereits vor, obwohl das Projekt noch andauert und der Inhalt bzw. die Struktur des Korpus noch nicht vollständig ausgewogen sind. Desiderata für weitere Anwendungen der Methoden und für die Erweiterung des Projektes schließen den Beitrag ab.

1 Einleitung

Außerparlamentarische politische Kommunikation stellt bezüglich der Datenerhebung eine besondere Herausforderung dar. Offiziell redigierte Transkripte stehen i. d. R. nicht zur Verfügung, und die Materialien sind nicht leicht zugänglich. Das Projekt Po.La.R. (*Political Language Repository*) zielt daher darauf ab, ein kollaboratives Korpus außerparlamentarischer Kommunikation in deutscher, italienischer, französischer und spanischer Sprache zu erstellen. Das Korpus sammelt orthographische Transkripte und steht für linguistische Analysen zur Verfügung. Das Projekt geht aus einer vorherigen Arbeit zum Rechtspopulismus [1] hervor. Das Po.La.R.-Korpus ist hingegen parteiübergreifend angesetzt und enthält keine spezifischen Bezüge zum Populismus.

Im Mittelpunkt des Projekts steht die politische außerparlamentarische Kommunikation, die eine relevante Rolle in politischen Wahlkämpfen sowie in politischen Ereignissen spielt. Ihre Erhebung ist jedoch mit einem großen Zeitaufwand verbunden, da die Materialien zunächst transkribiert werden müssen, um mithilfe linguistischer Instrumente untersucht werden zu können. Die Verfügbarkeit bereits aufbereiteter und zugänglicher Materialien ermöglicht es, den Zeitaufwand deutlich zu reduzieren, da man nicht jede Rede eigens orthographisch transkribieren muss. Teil des Korpus sind auch multimodale Annotationen, die mittels ELAN [2] bearbeitet werden und multimodale Analysen erlauben. Falls eine benötigte Quelle noch nicht im Korpus enthalten ist, kann sie integriert werden.

Obwohl bereits Projekte zu Datenbanken politischer Kommunikation vorhanden sind, befassen sich diese tendenziell entweder mit parlamentarischer Kommunikation [3] oder – in den meisten Fällen – mit einer einzelnen Gattung, wie z. B. TV-Debatten, Interviews usw., bzw. mit einer einzelnen Sprache [4]. Ein Korpus gesprochener politischer Kommunikation, das

[1] Dieser Beitrag entstammt von der Zusammenarbeit beider Autor:innen. Abschnitte 2 (*Methoden*) und 3 (*Erste Ergebnisse*) wurden von Marcella Palladino verfasst, während Abschnitte 1 (*Einleitung*) und 4 (*Ausblick*) von Vincenzo Gannuscio stammen.

Methoden der Sprachverarbeitung in die Politolinguistik einführt, wurde von Palladino [5] erstellt; dieses umfasst jedoch sowohl parlamentarische als auch außerparlamentarische Kommunikation. Ein Korpus, das ausschließlich vielfältige Formen außerparlamentarischer gesprochener politischer Kommunikation abdeckt, scheint bislang nicht zu existieren. Analysen politischer Sprache können jedoch unterschiedliche außerparlamentarische Gattungen umfassen. Hinzu kommt das Ziel, das Korpus *kollaborativ* aufzubauen, was bedeutet, dass Nutzer:innen aktiv zu seiner Erweiterung beitragen, und ihre eigenen bereits vorhandenen Materialien einbringen können.

Das Korpus umfasst Daten aus dem Zeitraum 2014–2025 und zielt damit vor allem auf die Erfassung aktueller, nicht historischer politischer Kommunikation. Der Umfang des Korpus ist möglichst breit konzipiert; synchronische oder diachronische Analysen werden jedoch nicht im Rahmen des Projektes selbst durchgeführt, sondern können von den Nutzer:innen eigenständig vorgenommen werden. Die Auswahl der Dateien richtet sich in erster Linie nach der Verfügbarkeit frei zugänglicher öffentlicher Quellen auf der Plattform YouTube. Aus diesem Grund können Beiträge derselben politischen Akteur:innen aus unterschiedlichen Jahren im Korpus vertreten sein, ohne dass dies ein eigenes Auswahlkriterium darstellt. Weitere Auswahlkriterien sind die parteiübergreifende Repräsentativität sowie das Ziel, möglichst umfangreiches Material für das Korpus zu sammeln.

2 Methoden

Die Dateien werden auf YouTube gesammelt und mittels KI-gestützter ASR-Systeme transkribiert. Das Korpus umfasst sowohl unkorrigierte ASR-Transkripte als auch manuell korrigierte Transkripte, wobei die Tokens jeweils separat berechnet werden. Unter *Tokens* versteht man im Projekt die Wörter, die in den Reden erscheinen. Diese Definition wird auf Hinblick des kollaborativen Charakters des Korpus verwendet und ermöglicht unter anderem korpuslinguistische Auswertungen wie Berechnung der Type-Token-Ratio.

Die im Korpus vertretenen Textgattungen sind Wahlkampfreden, Parteitagsreden, Pressekonferenzreden, Interviews und Talkshows. Als ASR-System kommt Whisper OpenAI [6] zum Einsatz; die Transkriptionskonventionen wurden eigens für dieses Projekt entwickelt. Die Reden werden zudem mit Abkürzungen versehen, die die progressive Reihenfolge des Transkriptionsprozesses kennzeichnen (z. B. bezeichnet CHR_2021_002 eine Rede von Tino Chrupalla, aus dem Jahr 2021, die als zweite Rede dieses Sprechers und dieses Jahres in das Korpus aufgenommen wurde). Eine Legende der Abkürzungen sowie der vollständigen Namen, der Sprache, der Partei, des Geschlechts und des Geburtsdatums der Redner:innen ist im Korpus verfügbar. Bisher wurden ausschließlich Dateien aus Deutschland, Frankreich, Spanien und Italien berücksichtigt. Für die zukünftige Erweiterung und Weiterentwicklung des Projektes ist vorgesehen, politische Kommunikation in denselben Sprachen auch aus weiteren Ländern in das Korpus aufzunehmen.

Die Videos werden aus YouTube in Audios konvertiert und mit ASR transkribiert. Bisher wurde Whisper OpenAI als ASR-System in Python [7] und mithilfe der App aTrain [8] angewandt. Die App aTrain erlaubt es, verschiedene Sprecher:innen in einem Audio zu erkennen, was für Redegattungen (z. B. Interviews), in denen eine Interaktion stattfindet, besonders hilfreich ist. In Python werden die Modelle Small, Medium und Large durch FasterWhisper verwendet, und für jede Datei wird notiert, welches Modell eingesetzt wird.

Das ausgewählte Output-Format ist aus zwei Gründen die TXT-Datei: Zum einen lassen sich TXT-Dateien leicht korrigieren bzw. bearbeiten, was die Aufgabe der manuellen Korrektur der von ASR erzeugten Transkripte erleichtert. Zum anderen ist das TXT-Format von den meisten Softwareprogrammen und Tools problemlos lesbar und kann bei Bedarf einfach in andere Formate konvertiert werden.

Neben den TXT-Transkripten enthält das Korpus auch EAF-Dateien, die mit der Software ELAN annotiert werden. Derzeit wird die multimodale Annotation deutscher Wahlkampfreden

erprobt, insbesondere im Hinblick auf relevante multimodale Analysekategorien. Diese Arbeit hat zugleich einen didaktischen Charakter, da die Annotationen zunächst von BA- und MA-Studierenden der Germanistik an der Universität Modena und Reggio Emilia erstellt werden. Anschließend werden sie überprüft, und die orthographischen Transkripte aus den TXT-Dateien werden als eigene Spur in ELAN integriert.

3 Erste Ergebnisse

Das Korpus umfasst 2.041.714 Tokens aus KI-gestützten ASR-Systemen und 181.078 Tokens aus korrigierten Transkripten (Stand: 20.10.2026). In Abb. 1 wird die Aufteilung der Tokens nach Sprachen gezeigt.

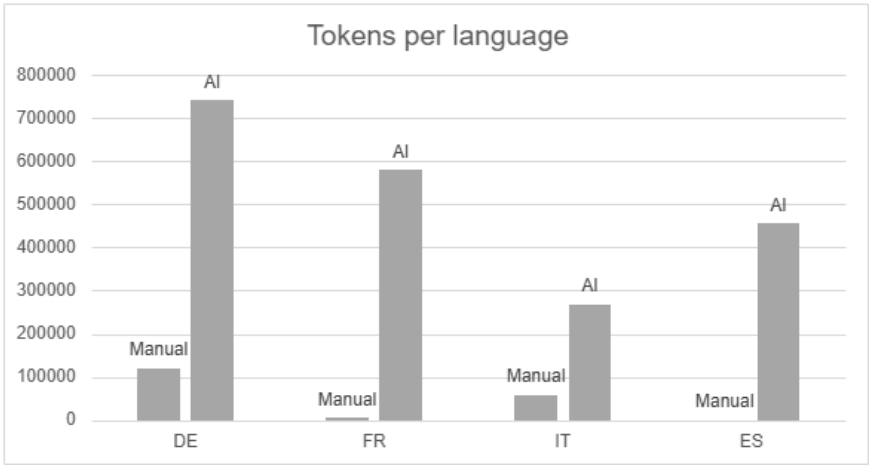

Abbildung 1 - Aufteilung der Tokens (Stand: 20.01.2026)

Die Säulen geben für jede Sprache die Anzahl der KI erzeugten Tokens sowie der anschließend manuell korrigierten Tokens an. Es handelt sich also nicht um zwei unterschiedliche Einheiten, sondern um dieselben Reden, die zunächst automatisch transkribiert und danach manuell überprüft wurden. *Manual* (manuell) bedeutet in diesem Zusammenhang, dass die Dateien von Personen korrigiert und gemäß den festgelegten Transkriptionskonventionen bearbeitet wurden. Keine der Dateien wurde ohne vorherige ASR-Transkription erstellt.

Abb. 2 zeigt die gesamte Dauer der Dateien je nach Sprache.

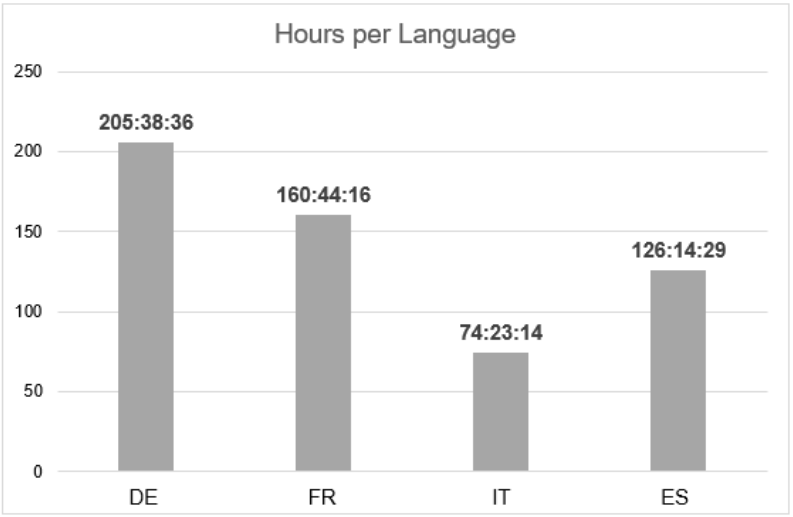

Abbildung 2 - Dauer der Dateien in Stunden:Minuten:Sekunden (Stand: 20.01.2026)

Das Korpus ist bislang noch nicht ausgewogen und die Datenerhebung wird stets weitergeführt, um repräsentative Abdeckung in Bezug auf Gattung, Geschlecht, Sprache, Partei und Zeitraum sicherzustellen. Die orthographischen Transkripte werden von Studierenden und Mitarbeiter:innen des Projektes korrigiert. Wie Abb.1 zeigt, ist die Anzahl der korrigierten Tokens pro Sprache jedoch noch nicht homogen. Während deutsche und italienische Transkripte bereits im Rahmen des vorherigen Projektes teilweise schon korrigiert wurden und Teil von Po.La.R. geworden sind, mussten französische und spanische Dateien zunächst erhoben werden. Im vorherigen Projekt war Spanisch noch nicht vorhanden, während französische Materialien großenteils schriftlich statt mündlich waren.

Zum Korpus gehören bisher korrigierte und unkorrigierte orthographische Transkripte, Videos und Audios (die nicht öffentlich zugänglich sind und ausschließlich für die ASR-Verarbeitung genutzt werden), frei zugängliche YouTube-Links und ELAN-annotierte Dateien von 11 deutschen Wahlkampfreden (Stand: 20.01.2026). TextGrid-Dateien sowie EXB-Dateien (diese wurden mit dem Softwareprogramm EXMARaLDA [9] in Partitur transkribiert) einiger Reden sind unsystematisch vorhanden und die Mehrheit der Reden verfügt bisher nicht darüber.

Nach Abschluss des Korpus werden die Transkripte, Annotationen sowie die YouTube-Links zu den Reden öffentlich zugänglich gemacht. Zusätzlich sollen weitere von Nutzer:innen bereitgestellte Materialien sowie bislang fehlende Dateiformate der vorhandenen Daten nach und nach integriert werden. YouTube ist eine frei zugängliche Plattform [10], die öffentliche Inhalte ohne verpflichtende Anmeldung bereitstellt. Da die dort verfügbaren Links jedoch häufig nur temporär bestehen und aus urheberrechtlichen Gründen lediglich auf die Quellen verwiesen werden darf, nicht aber die Audio- oder Videodateien selbst weitergegeben werden können, ergibt sich eine Herausforderung hinsichtlich der langfristigen Verfügbarkeit und nachhaltigen Nutzbarkeit des Korpus.

4 Ausblick

Dieser Beitrag soll den Ausgangspunkt für eine Diskussion über Methoden und Optionen bilden, die für das Po.La.R.-Korpus nützlich sein können. Das Projekt hat im Dezember 2024 angefangen und Kollaborationen mit Expert:innen der untersuchten Sprachen sind bereits im Gange. Das Interesse für eine weitere Diskussion gilt insbesondere alternativen Transkriptionssystemen zu Whisper OpenAI sowie den Grenzen des angewandten Verfahrens in Bezug auf Transkription und Datenerhebung.

Zudem werden derzeit Möglichkeiten für geeignete Plattformen geprüft, um den Zugriff auf das Korpus zu ermöglichen. Kollaborative Ansätze für weitere Projekte sind erwünscht, welche sich nicht auf die politische Sprache beschränken müssen. Das Projekt bietet eine Unterstützung in der Analyse der außerparlamentarischen Kommunikation, in der die Erhebung und Bearbeitung der Dateien nicht leicht erscheinen. Der entwickelte Ansatz kann jedoch auch für andere Disziplinen von Nutzen sein, in denen ähnliche Herausforderungen bestehen. Ein weiteres Desideratum besteht darin, das Projekt auf weitere Sprachen und Länder zu erweitern, damit das Korpus eine größere Repräsentativität (und Nutzbarkeit) haben kann.

Danksagung

Diese Arbeit wurde von Università di Modena e Reggio Emilia – Fondazione di Modena Projekt „CUP E93C24001970005 Beyond Parliament: AI-Enhanced Multilingual Corpus Using Innovative Methodology for Non-Institutional Political Speeches in German, French, Spanish and Italian“, Fondo di Ateneo per la ricerca Anno 2024 - Bando per il finanziamento di progetti di ricerca interdisciplinari finanziert.

Literatur

[1] Modena, S., M. Palladino, und V. Gannuscio: *A Multilingual Corpus of German, French and Italian Political Discourse: Goals and Methodological Challenges*. In: S. Grawunder (Hrsg.): *Elektronische Sprachsignalverarbeitung 2025. Tagungsband der 36. Konferenz*, S. 125 – 129. TUDpress, Dresden, 2025.

[2] Elan (Version 7.0): Nijmegen: Max-Planck-Institute for Psycholinguistics, The Language Archive, 2025. https://archive.mpi.nl/tla/elan (letzter Abruf 31.12.2025).

[3] Erjavec, T., M. Kopp, N. Ljubešić, et al.: *ParlaMint II: advancing comparable parliamentary corpora across Europe*. In *Language Resources & Evaluation*, 59, S. 2071 – 2102. 2025.

[4] Trotta, D., S. Tonelli, A. Palmero Aprosio, und A. Elia: *Annotation and Analysis of the PoliModal Corpus of Political Interviews*. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. 2019.

[5] Palladino, M.: *Politolinguistics through Spoken Language Processing: A Methodological Framework for German and Italian Political Speeches*. In: S. Grawunder (Hrsg.): *Elektronische Sprachsignalverarbeitung 2025. Tagungsband der 36. Konferenz*, S. 204 – 211. TUDpress, Dresden, 2025.

[6] Radford, A., J. Q. Kim, T. Xu, G. Brockman, C. McLeavey, und I. Sutskever: *Robust speech recognition via large-scale weak supervision*. In *Proc. of the 40th International Conference on Machine Learning*, S. 28492 – 28518. 2023.

[7] Python Software Foundation: Python 3.13.5, 2025. https://www.python.org/ (letzter Abruf 31.12.2025).

[8] Haberl, A., J. Fleiß, D. Kowald, und S. Thalmann: *Take the aTrain. Introducing an interface for the Accessible Transcription of Interviews*. In *Journal of Behavioral and Experimental Finance*, Vol. 41. 2024.

[9] Schmidt, T., und K. Wörner: *EXMARaLDA*. In *Handbook on Corpus Phonology*, S. 402 – 419. Oxford University Press, 2014. www.exmaralda.org (letzter Abruf 19.01.2026).

[10] Androutsopoulos, J., J. Tereick: *YouTube: Language and Discourse Practices in Participatory Culture*. In: T. Spilioti, A. Georgakopoulou (Hrsg.): *The Routledge Handbook of Language and Digital Communication*, S. 354 – 370. Routledge, Abingdon/New York, 2016.