Günther Wirsching (Hrsg.)
Elektronische Sprachsignalverarbeitung 2026
Tagungsband der 37. Konferenz
Eichstätt, 4.-6. März 2026

TUDpress

Studientexte zur Sprachkommunikation
Hg. von Rüdiger Hoffmann
ISSN 0940-6832
Bd. 113

Günther Wirsching (Hrsg.)

Elektronische Sprachsignalverarbeitung 2026

Tagungsband der 37. Konferenz

Eichstätt, 4.-6. März 2026

TUDpress

2026

Einbandfoto, Hintergrund: © Günther Wirsching
Einbandfoto Zentralmotiv: © Günther Wirsching

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

ISBN 978-3-95908-834-3

D-01309 Dresden
Tel.: +49 351 4721463
http://www.tudpress.de

TUDpress ist ein Imprint von Thelem

Gesetzt von den Herausgebern.
Printed in Germany.

Vorwort

Liebe Freundinnen und Freunde der ESSV, liebe Kolleginnen und Kollegen,

Veranstaltungsort der 37. Konferenz *Elektronische Sprachsignalverarbeitung* 2026 ist die Katholische Universität Eichstätt-Ingolstadt, wo schon die 26. ESSV 2015 stattfand.

Auch diesmal ist die ESSV thematisch breit gefächert, von der Sprachsignalanalyse über verschiedene Techniken zur Sprachanalyse und Methoden der Sprachsynthese bis zu Dialogsystemen und kognitiven Aspekten des Sprachverstehens, wobei auch Verbindungen zur künstlichen Intelligenz Berücksichtigung finden.

Das Programm der ESSV 2026 enthält vier Hauptvorträge und neunzehn Fachvorträge sowie mehrere Poster- oder Show & Tell-Beiträge. Wir freuen uns, dass wir Hauptvorträge zu den Bereichen Geschichte, Anglistik und Anwendungen im Bereich Medizin gewinnen konnten. Ein weiterer Grund zur Freude ist die hohe Qualität der weiteren fachlichen Beiträge, die auf der Konferenz präsentiert werden.

Mein persönlicher Dank gilt dem Organisator der ESSV 2025, Sven Grawunder, sowie Kollegen aus dem ESSV-Förderverein, die mich bei der Organisation der Konferenz stark unterstützt haben.

Günther Wirsching
Eichstätt, Januar 2026

Inhaltsverzeichnis

Speech Analysis I

Speech Synthesis

Speech Analysis II

Voice, Language and Cognition

Posters

Show & Tell

ZUR TRANSKRIPTION MÜNDLICHER PHÄNOMENE IN DER POLITISCHEN SPRACHE

Marcella Palladino

Università degli Studi di Modena e Reggio Emilia
marcella.palladino@unimore.it

Kurzfassung: Dieser Beitrag befasst sich mit ASR-Transkripten außerparlamentarischer politischer gesprochener Kommunikation auf Deutsch, die mit dem Zweck erstellt werden, linguistische Analysen durchzuführen. Um gesprochene Dateien zu transkribieren, stellen ASR-Systeme vor allem im Falle großer Datenmengen oder langer Reden eine wichtige Ressource dar. Es ist bekannt, dass viele ASR-Systeme die Sprache tendenziell normalisieren, und in dieser Studie wird dargestellt und diskutiert, welche Probleme für linguistische Analysen entstehen können. Der Fokus liegt dabei spezifisch auf Phänomenen der Mündlichkeit wie Interjektionen und Füllwörtern, die in vielen Fällen keine eindeutig identifizierbaren lexikalischen Einheiten darstellen, aber trotzdem unerlässlicher Teil der Kommunikation sind. Diese Elemente werden von ASR-Systemen oft nicht erkannt, vor allem, wenn man öffentlich zugängliche Instrumente verwendet, wie sie in diesem Beitrag eingesetzt werden. Diese mündlichen Elemente besitzen allerdings in der politischen Sprache eine große Relevanz und sind für linguistische Analysen von großem Interesse. Der Beitrag veranschaulicht diese Problematik anhand von Beispielen und enthält die Diskussion sowohl über die Vorteile als auch die Herausforderungen ausgewählter öffentlich verfügbarer ASR-Systeme für die genannten Forschungszwecke. Abschließend wird die Hoffnung geäußert, dass die Dateien der außerparlamentarischen Kommunikation als Ressource für öffentliches aber gezieltes Modelltraining betrachtet werden können.

1 Hintergrund und Ziele der Studie

ASR-Systeme (*Automatic Speech Recognition*) ermöglichen es heutzutage, große Mengen von mündlichen Dateien in kurzer Zeit zu transkribieren. Allerdings wird die Sprache von diesen Systemen i. d. R. normalisiert und eine manuelle Überprüfung der Transkripte erscheint notwendig, um Verbatim-Transkripte zu erhalten.

In diesem Beitrag werden unter dem Begriff *mündliche Phänomene* Interjektionen und Füllwörter sowie weitere vergleichbare Elemente verstanden, die sowohl lautliche als auch lexikalische Einheiten der deutschen Sprache sein können. Die Analyse fokussiert sich dabei insbesondere auf die lautlichen Einheiten. Der Ausdruck *mündliche Phänomene* wird ausgewählt, da kein geeigneter alternativer Oberbegriff gefunden werden konnte, der nicht spezifisch auf bestimmte Funktionen dieser Phänomene hinweist. Die Begriffe *Interjektionen*, *Füllwörter* usw. sind bezüglich ihrer Definitionen etwas umstritten. Beispiel dafür kann die Definition von Dingemanse [1] sein, laut der Interjektionen als "words or short phrases that typically constitute an utterance in a larger interactive sequence" gelten. Verzögerungslaute[1] gelten beispielsweise nicht (immer) als Interjektionen, da sie nicht unbedingt als eigenständige Äußerungen funktionieren und sie auch in monologischen Situationen vorkommen können. Einige Formen könnten aber z. B. sowohl als Interjektionen als auch als Füllwörter je nach dem Kontext gelten. Oft

[1] In diesem Beitrag werden die Termini *Verzögerungslaute* und *Füllwörter* weitgehend synonym verwendet. Laute wie *ähm* können Pausen *füllen*; zugleich wäre jedoch eine genauere Definition des Begriffs *Wort* erforderlich. Diese begriffliche Erklärung wird hier nicht vertieft, wobei festzuhalten ist, dass der Begriff *Füllwörter* m. E. kontrovers bleibt.

wird das Wort *Interjektionen* auch als Oberbegriff verwendet, um Partikel, Wörter und Laute einzubeziehen, die, so Duden [2], „keine syntaktische Funktion“ besitzen. In diesem Beitrag wird keine Unterteilung dieser Begriffe gemacht bzw. begründet, jedoch bleibt das Thema für künftige spezifische Analyse interessant.

Viele Studien sowohl in der Linguistik als auch in der ASR-Entwicklung befassen sich mit Interjektionen und lautlichen Phänomenen sowie mit ihrer Transkription. So erforschen etwa Dingemanse und Liesenfeld [3] Interjektionen und weisen darauf hin, dass sie in Analysen leicht zu übersehen sind, da sie sowohl in der Transkription von Dateien als auch in der Erstellung von Korpora gesprochener Sprache unterrepräsentiert bleiben. Gorisch und Schmidt [4] heben hervor, dass die für Linguist:innen relevanten Transkriptionskriterien nicht mit jenen übereinstimmen, die ASR-Entwickler:innen zur Bewertung der technischen Qualität berücksichtigen. Sie betonen zudem, dass *disfluencies* sowie auch die Grammatik durch ASR-Systeme normalisiert werden können, was eine Hürde für die Erstellung linguistischer Korpora ist. Sie stellen zusätzlich in Frage, ob Parameter wie WER (*Word Error Rate*) für den Vergleich zwischen manuell korrigierten ASR-Verbatim-Transkripten und manuellen Verbatim-Transkripten aufgrund der von ASR nicht-transkribierten mündlichen Phänomene geeignet sein können.

Darüber hinaus gibt es Elemente wie *ach so,* die ebenfalls als Interjektionen gelten, die von ASR-Systemen mit höherer Wahrscheinlichkeit als Laute wie *ähm* erkannt werden könnten, da sie als Elemente der deutschen Sprache gelten. Pausen stellen hingegen kein großes Hindernis für die Transkription dar, weil die ASR-Systeme i. d. R. es auch ermöglichen, Zeitmarkierungen einzusetzen, was die Präzision bei der Pausenannotation vereinfachen und verbessern kann.

Mündliche Phänomene wie Interjektionen, Verzögerungslaute, Pausen usw. prägen die gesprochene Sprache. Die parlamentarische Kommunikation ist i. d. R. aufgrund ihrer vorherigen Vorbereitung stärker formalisiert als die außerparlamentarische Kommunikation, sodass solche Phänomene seltener vorkommen. Außerparlamentarische gesprochene Kommunikation kann hingegen als authentischer gelten, was die mündlichen Phänomene angeht (solange die Reden nicht vollständig vorformuliert bzw. vorgelesen werden). In der politischen Kommunikation gewinnen diese Phänomene auch an Bedeutung, weil sie zur Wirksamkeit sowie zur Persuasion einer Rede beitragen können [5] [6]. Studien haben sich mit den akustischen Merkmalen sowie mit der Prosodie politischer Kommunikation beschäftigt [7] [8], aber Phänomene wie Interjektionen und Füllwörter werden meist lediglich signalisiert bzw. quantifiziert. Ihre Funktionen in der politischen außerparlamentarischen Kommunikation können jedoch aus vielen Gründen relevant sein. Z. B. gewinnen sie im Rahmen des Machtverhältnisses der Politiker:innen mit dem Publikum bzw. den Gesprächspartner:innen an Relevanz und es wäre wünschenswert, untersuchbare Materialien zur Verfügung zu haben. Zudem können diese Elemente vielsagend sein, wenn man die Kommunikation unterschiedlicher Politiker:innen vergleichen mag oder die Reden ausgewählter Politiker:innen diachronisch analysieren will.

In diesem Beitrag wird untersucht, wie mündliche Phänomene wie Interjektionen und Füllwörter von den ASR-Systemen transkribiert werden und welche Herausforderungen entstehen, wenn man die Transkripte für linguistische Analysen verwenden mag. Des Weiteren wird die linguistische Relevanz dieser mündlichen Phänomene bei der Analyse politischer Sprache hervorgehoben.

Die Quelle dieser Studie ist das Po.La.R.-Korpus (*Political Language Repository*) der Universität Modena und Reggio Emilia. Dieses Korpus ist multilingual und enthält transkribierte Dateien außerparlamentarischer Kommunikation (d. h. mehrerer Gattungen) auf Deutsch, Französisch, Italienisch und Spanisch. Für diese Studie werden nur deutsche Dateien berücksichtigt. Der deutsche Teil des Korpus ist bisher die strukturierteste Sektion und für die vorliegende Analyse liegen bereits manuell korrigierte Dateien vor. Die Transkriptionskonventionen wurden für das Projekt entwickelt und folgen keinem bestehenden System. Sie werden in der manuellen Überprüfung der Transkripte eingesetzt.

2 Methoden

Die Transkripte des Po.La.R.-Korpus (im TXT-Format) basieren auf YouTube-Videos, die in WAV-Audios konvertiert und mit Whisper OpenAI [9] transkribiert werden. Whisper wird in Python über FasterWhisper sowie in der App aTrain [10] angewandt. Die Modelle Small, Medium und Large werden in Python verwendet, während für aTrain das Modell Large angewandt wird und die automatische Erkennung der verschiedenen Sprecher:innen ausgewählt wird. Die ASR-Transkripte werden anschließend manuell korrigiert.

Für das Projekt wird FasterWhisper verwendet, da es ein schnelles Transkriptionsverfahren bietet, das den Zielen der Datenerhebung für das Po.La.R.-Korpus entspricht (vgl. Palladino und Gannuscio in diesem Band). Die Transkripte werden entweder in Python oder mithilfe der App aTrain generiert, die ebenso FasterWhisper verwendet. FasterWhisper scheint also für die Erstellung eines breiten Korpus und für den abwechselnden Einsatz von Python und aTrain passend. Am Korpus arbeiten Mitarbeiter:innen mit unterschiedlichen Forschungshintergründen und die beiden Optionen (Python oder aTrain) ermöglichen es auch denjenigen ohne Programmierkenntnisse, zur Transkription der Dateien beizutragen. Der Grund, warum in Python die Modelle Small, Medium und Large verwendet werden, hat mit der technischen Ausstattung des Projektes zu tun. Zuerst stand ein Rechner zu Verfügung, der nur das Modell Small unterstützen konnte und danach wurde ein Rechner gekauft, der hingegen auch die anderen Modelle unterstützt. Das, was als eine methodische Grenze gesehen werden kann, ist tatsächlich eine Möglichkeit, um unterschiedliche Modelle von Whisper OpenAI für die Ziele des Projektes zu verwenden und zu testen.

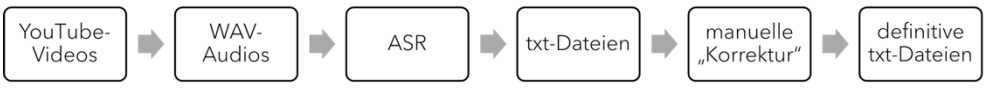

Abbildung 1 - Workflow der Transkription

In Abbildung 1 wird der Workflow der Transkription dargestellt. Das Wort *Korrektur* wird benutzt, um auf die Überprüfung der Transkripte hinzuweisen, bei der eventuelle von der ASR beibehaltene Fehler der Sprecher:innen nicht korrigiert werden. Zudem werden Fehler der Sprecher:innen wieder eingefügt, falls sie von der ASR normalisiert werden. Bei der manuellen Korrektur wird auch überprüft, ob die oben genannten mündlichen Phänomene vom System transkribiert werden. Falls dies nicht der Fall ist, werden sie je nach den Transkriptionskonventionen hinzugefügt.

Whisper-Implementierungen für *speech disfluencies detection* sind vorhanden. Uns ist z. B. CrisperWhisper [11] bekannt, das die Erkennung von *speech disfluencies* erlaubt: “Since our model transcribes verbatim, we also detect and segment other disfluency types, such as repetitions, false starts or partial words”. Allerdings wurde CrisperWhisper für das Po.La.R.-Korpus nicht eingesetzt, da eine der Voraussetzungen für das Po.La.R.-Projekt die Verwendung öffentlich zugänglicher Quellen und Instrumente ist. Als Plattform für die Datenerhebung wurde daher YouTube gewählt, weil es eine freie offene Quelle ist, die keine Anmeldung benötigt. CrisperWhisper benötigt hingegen eine Anmeldung und entspricht damit nicht den Zielsetzungen des Projektes. Auch andere Weiterimplementierungen wie z. B. [12] [13], die vor allem für klinische Analysen konzipiert wurden, ermöglichen die Erkennung von *speech disfluencies*. Diese Implementierungen wurden jedoch für spezifische klinische Datentypen entwickelt. Die außerparlamentarischen Aufnahmen, die uns zu Verfügung stehen, haben i. d. R. keine optimale

Audioqualität und wurden nicht eigens für die Erstellung eines linguistischen Korpus produziert, sondern sie stammen aus öffentlich zugänglichen YouTube-Videos.

3 Analyse

In diesem Abschnitt wird ein politisches Interview mittels der zwei Optionen des Po.La.R.-Korpus transkribiert und manuell korrigiert. Die Ergebnisse werden aus linguistischer Perspektive verglichen und kommentiert. Das Interview ist aus einer Talkshow des Jahres 2025 entnommen, an der eine Politikerin und ein Moderator teilnehmen. Das Video[2] ist in YouTube öffentlich zugänglich und ist besonders interessant, weil das Verhältnis zwischen den Gesprächsteilnehmer:innen nicht entspannt aussieht. Der Moderator stellt Fragen, die eine Haltung erkennen lassen, die der Position der Politikerin entgegengesetzt ist; diese beklagt ihrerseits mehrmals, nicht ausreden zu dürfen. In dieser Situation kann man die Entwicklungen sowie die Verteidigung eines Machtverhältnisses im Laufe der Interaktion beobachten. Dazu tragen mündliche Phänomene bei. Z. B. versucht die Politikerin so wenig wie möglich lautlos zu pausieren, damit sie ihr Rederecht nicht verliert. Dabei kommen Laute vor, die mal als Interjektionen und mal als Füllwörter (im Sinne, dass sie die Pausen *füllen*) wirken. Aus der Perspektive der linguistischen Analyse sind diese Phänomene sehr interessant, weil sie die politische Kommunikation maßgeblich prägen und sowohl implizit als auch explizit die Intentionen sowie die Argumentations- und Verteidigungsstrategien der Politikerin signalisieren. Sie sind daher Bestandteile, die in Transkripten behalten werden sollen.

Die Analyse wird mit einem qualitativen Ansatz durchgeführt, d. h., dass einzelne Beispiele gezeigt und kommentiert werden. Bisher ist aus zwei Gründen keine quantitative Analyse vorhanden. Einerseits dauert das Projekt noch an und selbst die Transkripte, die bereits zur Verfügung stehen, müssen ein zweites Mal überprüft werden, bevor das Korpus veröffentlicht wird. Andererseits versteht sich diese Studie eher als eine Illustrierung der Herausforderungen, die hoffentlich weiter diskutiert und möglicherweise für die weitere Implementierung des Korpus und der Methoden gelöst werden können.

Das ausgewählte Interview befindet sich mit der Dateikennzeichnung WEI_2025_003 im Po.La.R.-Korpus. Für die vorliegende Studie werden drei im Po.La.R.-Korpus vorhandene Transkriptionen berücksichtigt: die von Whisper Small in Python, die von Whisper Large aus der App aTrain und die von mir manuell korrigierte Version. Small wurde in Python zuerst für das Projekt aufgrund der limitierten technischen Ausstattung eingesetzt, zudem ist das Modell bei der Transkription schneller. Small dürfte auch weniger als Large normalisiert sein [14] und daher könnte man erwarten, dass das vorteilhaft für die Verbatim-Transkription ist.

Ausgewählte Beispiele der manuell korrigierten Version werden gezeigt und es wird dargestellt, ob und in welcher Form mündliche Phänomene in den ASR-transkribierten Versionen vorkommen. Ihre Relevanz in der Analyse wird ebenso erläutert. Die berücksichtigten Phänomene werden kursiv markiert.

Tabelle 1 - Nummerierte Beispiele von der manuell korrigierten Version

(1)	ich will es Ihnen gerade erklären *also* zunächst einmal gilt das Lohnabstandsgebot *ähm* von *ähm* nicht
(2)	ich will es Ihnen einfach nur mal erklären vielen Dank dass Sie mich ausreden lassen und *ähm* das Bürgergeld *ähm* […] ziemlich einfach *ähm* ziemlich einfaches Prinzip
(3)	*also ähm* sie machen eigentlich überhaupt gar keinen Druck sondern es ist recht einfach […] und um es hier ganz klar zu sagen für ausländische *ähm ähm ähm ja* Bürger
(4)	aber plötzlich Auszahlung bekommen dann kippt ihr Sozialversicherungssystem wie in Deutschland *ne so ne* die Arbeitslosen

[2] https://www.youtube.com/watch?v=kEqDks-h-qM (letzter Abruf 16.01.2026).

(5)	*nein nein nein* Sie vermischen die Kriterien
(6)	*nein* ich möchte es Ihnen einfach erklären Sie vermischen die Kriterien das Kriterium ist und ich habe es Ihnen eben auch schon mal zweimal erklärt *ähm* darum macht es das alles so schwierig *ähm* darum nochmal ganz einfach damit es hoffentlich verstanden wird es geht
(7)	ich sage Ihnen ganz klar die ganzen Millionen von Leuten die in unser Sozialversicherungssystem eingewandert sind die nicht eine Minute gearbeitet haben und hier Sozialleistung auf dem Rücken der arbeitenden Bevölkerung *äh äh* beziehen
(8)	<S0> *heh* <S1> jetzt reden wir von der indischen IT-Spezialistin <S0> *ja*
(9)	die Kriterien *ah* die die die ganzen Spezialisten die kommen aus ganz anderen Gründen nicht
(10)	nicht mehr kommt *so* aber nehmen wir mal den hypothetischen Fall

In Tabelle 1 befinden sich Beispiele der fokussierten mündlichen Phänomene. Auch Interjektionen und Füllwörter, die lexikalisierte Einheiten der deutschen Sprache sind, wurden einbezogen, um ihre Transkription mit der von Lauteinheiten zu vergleichen. In den Beispielen stehen auch die Po.La.R.-Transkriptionskonventionen sowie das Zeichen […], das von den Beispielen entfernte Wörter der gleichen Einheit ersetzt. <S0> bezeichnet die Politikerin und <S1> den Moderator. Es wurde versucht, Beispiele auszuwählen, die hauptsächlich die Wörter der Politikerin enthalten, damit man die Funktion der mündlichen Phänomene in Bezug auf die Sprache der politischen Vertreterin kommentieren kann. Sicher ist auch die Interaktion zwischen den Gesprächsteilnehmer:innen für weitere Analysen interessant. In Tabelle 2 werden die Versionen der ASR-Transkription gezeigt.

Tabelle 2 - Vergleich der ASR-Versionen

	Whisper Small	**Whisper Large**
(1)	*also* transkribiert	*also* transkribiert
(2)	nicht transkribiert	nicht transkribiert
(3)	*also* transkribiert	*also* transkribiert
(4)	nicht transkribiert	nicht transkribiert
(5)	transkribiert	transkribiert
(6)	*nein* transkribiert	*nein* transkribiert
(7)	nicht transkribiert	nicht transkribiert
(8)	nicht transkribiert	nicht transkribiert
(9)	nicht transkribiert	transkribiert
(10)	nicht transkribiert	transkribiert

Ein weiteres interessantes Beispiel ist der Teil der Interaktion, in dem die Politikerin lacht:

(11) müssen Sie das Programm gucken es ganz genau sich durchlesen und dementsprechend habe ich diese Frage jetzt ausreichend beantwortet <she laughs>

Sie lacht und sagt, dass sie die Frage des Moderators schon beantwortet habe. Dadurch impliziert sie, dass der Moderator ihre Antwort nicht verstanden hat und die Frage unangemessener Weise wiederholt. Das Lachen kommt in der Mitte des Wortes *beantwortet*, aber kein ASR-System erkennt einen Abbruch im Wort.

Wie man an den Beispielen erkennen kann, besitzen die übertragenen mündlichen Phänomene unterschiedliche Funktionen. Die Funktionen, die sie haben, sind aus linguistischer Perspektive interessant, da sie vieles über die Verwendung der Sprache und die Positionen der Politikerin sagen. Besonders interessant ist Beispiel (3). Der Teil „ausländische ähm ähm ähm

ja Bürger“ wird von keinem System transkribiert, jedoch sind sowohl *ähm* als auch die Interjektion *ja* in der Aussage relevant. *Ähm* könnte als Füllwort gelten oder es könnte strategisch eingesetzt werden, also nicht um lediglich eine Pause zu füllen, sondern als diskurspragmatisches Element: Leute, die aus anderen Ländern kommen, als Bürger zu bezeichnen, könnte nicht dem entsprechen, was die Politikerin tatsächlich sagen möchte. Durch das Füllwort könnte sie so tun, als wolle sie das richtige Wort finden. Die Interjektion *ja* könnte signalisieren, dass ihr das Wort eingefallen ist. Diese Verwendung von *ähm* und *ja* erweist sich womöglich als strategisch, weil diese Verzögerung gleichzeitig den Zweifel wiedergibt, ob man diese Leute *Bürger* bezeichnen sollte, und daher, ob sie zu Deutschland gehören.

Beispiel (8) ist aus linguistischer Perspektive ebenfalls sehr interessant. Die Politikerin antwortet mit einem irritierten lautlichen Seufzer und danach mit der Interjektion *ja*. Auf diese Weise signalisiert sie ihre Irritation und außerdem könnte sie implizieren, dass die Wörter des Moderators unnütz sind – entweder eine falsche Aussage oder eine Wiederholung. Sie äußert tatsächlich im Gespräch ihre Irritation darüber, auf die gleichen Fragen mehrmals antworten zu müssen, als ob der Moderator die Antworten nicht richtig verstehen würde. Im Machtverhältnis und linguistischen Profil bzw. in der Analyse der Politikerin in diesem Interview sind diese Elemente bedeutungsvoll. Vor allem am Ende des Interviews versucht die Politikerin, das Rederecht nicht zu verlieren, und zu diesem Zweck werden Füllwörter sowie Interjektionen häufiger eingesetzt. Beispiel dafür ist (9), in dem die Politikerin einen Satz anfängt, aber danach das Thema wechselt, nachdem sie die Interjektion *ah* verwendet hat. Auch im Beispiel (10) erkennt man beim Einsatz der Interjektion *so* den Versuch, weiterreden zu können und den Sprecherwechsel zu meiden.

4 Ergebnisse und Diskussion

In der Analyse sieht man deutlich, dass die Phänomene der Mündlichkeit, die als eigenständige Wörter gelten (z. B. *nein*, *also* usw.), manchmal transkribiert werden, vor allem von Whisper Large in der App aTrain. Whisper Large in der App aTrain hat ebenfalls die Interjektion *ah* im Beispiel (9) transkribiert. Andere lautliche Elemente wurden aber von beiden Modellen nicht wiedergegeben bzw. erkannt. Wie man in der Analyse merken konnte, sind die fokussierten mündlichen Phänomene für linguistische Forschungszwecke besonders wichtig und es wäre notwendig, sie in Transkripten zu haben. Die ASR-Transkription ist ein guter Weg, um eine große Datenmenge transkribieren zu können, aber die Transkripte müssen überprüft werden, um die Anwesenheit der mündlichen Phänomene zu kontrollieren bzw. sie hinzufügen zu können. In diesem Fall scheint FasterWhisper mindestens ein Instrument zu sein, um die Transkripte so schnell wie möglich zu produzieren, bevor sie überprüft werden.

Der Vergleich wurde nur anhand einer Datei gezeigt und qualitativ durchgeführt, daher sind die Ergebnisse nicht zu verallgemeinern. Allerdings konnte man anhand dieses Beispiels feststellen, wie herausfordernd es ist, die ASR-Transkription an linguistische Analysen anzupassen. Alles von Anfang an manuell zu transkribieren erscheint ebenfalls keine gute Lösung, insbesondere wenn man die Länge einiger Formen außerparlamentarischer Kommunikation berücksichtigt.

Wie oben erwähnt, könnten Implementierungen wie CrisperWhisper eine mögliche Lösung sein, um mündliche Phänomene wie Interjektionen, Füllwörter und andere *disfluencies* direkt transkribieren zu können, aber es erfordert eine Anmeldung, d. h., dass es nicht öffentlich verfügbar ist. Die Dateien des Po.La.R.-Korpus sind keine selbst durchgeführten Aufnahmen und *gated* Ressourcen stehen im Prinzip gegen die Voraussetzungen bzw. die Ziele eines kollaborativen Korpus wie Po.La.R. Methodisch würde die Verwendung von CrisperWhisper wahrscheinlich die Transkription der mündlichen Phänomene verbessern – aber nicht lösen, da die WER und die IER (*Insertion Error Rate*) nicht null sind und die Transkripte sowieso überprüft werden müssen. Methodisch ist es aber angemessen, es nicht zu verwenden, wenn die Nutzungsbedingungen gegen die Voraussetzungen des Projektes sind. Wünschenswert wäre es,

wenn öffentlich zugängliche Whisper-Implementierungen entwickelt würden, die *speech disfluencies* erkennen und für die Erstellung von Verbatim-Transkripten zwecks linguistischer Analysen geeignet sind.

Andere ASR-Systeme als Whisper könnten möglicherweise ebenso geeignete Lösungen bieten und Ziel dieses Beitrags ist es auch, zu dieser Diskussion beizutragen. Die Dateien des Po.La.R.-Korpus könnten bspw. als Trainingsdaten gelten und die mögliche Implementierung könnte die spezifischen Merkmale der Dateien (z. B. die Audioqualität, die Art von *speech disfluencies*, die öffentliche Natur der Dateien usw.) berücksichtigen. Eine auf diese spezifischen Merkmale zugeschnittene Anpassung würde voraussichtlich auch die Erkennungsrate der genannten mündlichen Phänomene erhöhen.

5 Fazit

Als Fazit kann die Relevanz der linguistischen Analyse mündlicher Phänomene wie Interjektionen und Füllwörter in der außerparlamentarischen politischen gesprochenen Kommunikation noch einmal hervorgehoben werden. Gleichzeitig zeigt sich die Notwendigkeit, angemessene Verfahren für ihre zuverlässige Transkription finden zu können, die auch für eine große Menge von Dateien geeignet erscheinen. Genau aus diesem Grund kann die ASR als vielversprechendes Instrument für die Erstellung linguistischer Korpora gelten, weil sie erlaubt, lange und oft schwer zu transkribierende Reden zur Verfügung zu stellen. Diese Transkripte gelten aber i. d. R. nicht als Endprodukt des Verfahrens, sondern benötigen eine Überprüfung. Diese Überprüfung könnte weniger zeitaufwendig sein, falls öffentlich zugängliche ASR-Systeme die genannten mündlichen Phänomene mindestens teilweise (zuverlässig) wiedergeben könnten.

Die Analyse der außerparlamentarischen Kommunikation geht im Projekt auch einen Schritt weiter, indem die Dateien multimodal annotiert werden. Dies ist auch im Fall von den genannten mündlichen Phänomenen relevant, da sie üblicherweise von nicht-mündlichen Phänomenen entweder begleitet oder ersetzt werden können (z. B. Augenbewegung, Stirnrunzeln usw.).

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] DINGEMANSE, M.: *Interjections at the Heart of Language*. In *Annual Review of Linguistics*, 10, S. 257 – 277. 2024.

[2] „Interjektionen“ auf Duden Sprachwissen online. URL: https://www.duden.de/sprachwissen/fuer-lernende/wortarten/interjektionen (letzter Abruf 14.01.2026).

[3] DINGEMANSE, M., und A. LIESENFELD: *From Text to Talk: Harnessing Conversational Corpo-ra for Humane and Diversity-Aware Language Technology*. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, Bd. 1, S. 5614 – 5633. Association for Computational Linguistics, Dublin, 2022.

[4] GORISCH, J., und T. SCHMIDT, T.: *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)*, S. 6564 – 6574. ELRA and

ICCL, Turin, 2024.
[5] ROSENBERG, A., und J. HIRSCHBERG: *Acoustic/Prosodic and Lexical Correlates of Charismatic Speech*. In *Interspeech2005*, S. 513 – 516. Lissabon, 2005.
[6] BRAGA, D., und M. A. MARQUES: *The pragmatics of prosodic features in the political debate*. In *Proc. Speech Prosody 2004*, S. 321 – 324. Nara, 2004.
[7] BURGEMEISTER, M., und W. F. SENDLMEIER: *Politische Sprechwirkungsforschung – ein Vergleich zwischen Alice Weidel, Annegret Kramp-Karrenbauer und Andrea Nahles*. In: R. W. WAGNER (Hrsg.): *Sprechen*, 69 (2020-1), S. 15 – 22. Verlag für Sprechwissenschaft und Kommunikationspädagogik, Heidelberg. 2020
[8] WEISS, B.: *Prosodic Elements of a Political Speech and its Effects on Listeners*. In *Proc. of the 10th International Conference SPEECH and COMPUTER 2005*, S. 127 – 130. Patras, 2005.
[9] 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.
[10] 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*. 2024.
[11] WAGNER, L., B. THALLINGER, und M. ZUSAG: *CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions*. In *Interspeech 2024*, S. 1265 – 1269. Kos, 2024.
[12] ROMANA, A., K. KOISHIDA, und E. M. PROVOST: *Automatic Disfluency Detection from Untranscribed Speech*. 2023. https://doi.org/10.48550/arXiv.2311.00867 (letzter Abruf 13.01.2026).
[13] AKINRINTOYO, E., N. ABDELHALIM, und N. SALOMONS: *WhisperD: Dementia Speech Recognition and Filler Word Detection with Whisper*. In *Interspeech 2025*, S. 1413 – 1417. Rotterdam, 2025.
[14] FEDOTOVA, A., A. FERRARESI, M. MILIČEVIĆ PETROVIĆ, und A. BARRÓN-CEDEÑO: *Constructing a Multimodal, Multilingual Translation and Interpreting Corpus: A Modular Pipeline and an Evaluation of ASR for Verbatim Transcription*. In *Proc. of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, S. 349 – 355. Pisa, 2024.