School of Engineering

Centre for Artificial Intelligence

Bachelorarbeit Informatik

Empirische Qualitätsbewertung für synthetische schweizerdeutsche Dialekte

Autoren		Fabio Huber Achille Martin Hünenberger
Hauptbetreuung		Jan Milan Deriu
Nebenbetreuung		Mark Cieliebak
Datum		05.06.2026

Zusammenfassung

Moderne Text-to-Speech-Systeme (TTS) erzielen bei der Zero-Shot-Voice-Adaptation fundamentale Fortschritte, stehen bei Low-Resource-Sprachen wie Schweizerdeutsch jedoch weiterhin vor komplexen Herausforderungen. Zwar existieren vielversprechende Synthesemodelle, die standarddeutsche Texte in sieben regionale Schweizer Dialekte synthetisieren können, es fehlt jedoch an einer methodisch fundierten Validierung der generierten Stimuli durch menschliche Annotationen.

Zur Schliessung dieser empirischen Forschungslücke entwickelt die vorliegende Arbeit eine systematische Evaluationsmethodik. Deren Kern bilden eine dedizierte Annotationsplattform mit integrierter Kalibrierung der individuellen Dialekterkennungskompetenz und eine automatisierte Pipeline zur Stimuligenerierung. Um der starken Subjektivität menschlicher Bewertungen zu begegnen, werden die erhobenen Daten mittels hierarchischer bayesscher ordinaler logistischer Regressionsmodelle ausgewertet, welche die Isolierung und mathematische Modellierung stimulusspezifischer sowie personenspezifischer Effekte ermöglichen.

Die durchgeführte Studie mit 168 Teilnehmenden resultierte in einem Datensatz von 11’365 Annotationen. Die statistische Auswertung zeigt, dass der personenspezifische Bias und die Annotationssicherheit das Bewertungsverhalten dominieren. Bei den stimulusspezifischen Parametern weist der Satz den weitaus grössten Einfluss auf, während die Referenzstimme und der Dialekt eine untergeordnete Rolle spielen. Die Arbeit leistet somit einen wesentlichen Beitrag zur Systematisierung der Dialekt-TTS-Verifikation und zeigt auf, dass demografisch ausgewogenere Stichproben die Aussagekraft der Methodik künftig noch weiter steigern können.

Abstract

Modern text-to-speech (TTS) systems are making significant progress in zero-shot voice adaptation, but they still face complex challenges when it comes to low-resource languages such as Swiss German. Although there exist promising synthesis models which can synthesise standard German texts into seven regional Swiss dialects, there is a lack of methodologically sound validation of the generated recordings through human annotation.

To address this gap in empirical research, this study develops a systematic evaluation methodology. At its core are a dedicated annotation platform with integrated calibration of individual dialect recognition ability and an automated pipeline for stimulus generation. To counteract the high degree of subjectivity in human assessments, the collected data is analysed using hierarchical Bayesian ordinal logistic regression models. These enable the isolation and mathematical modelling of recording-specific and participant-specific effects.

The study, which involved 168 participants, resulted in a dataset of 11,365 annotations. Statistical analysis shows that participant-specific bias and annotation confidence have a decisive influence on rating behaviour. Among the recording-specific parameters, the sentence has by far the greatest influence, whilst the reference voice and dialect play a secondary role. The study thus makes a significant contribution to the systematisation of dialect TTS verification and demonstrates that demographically more balanced samples could further enhance the validity of the methodology in the future.

Inhaltsverzeichnis

1 Einleitung 1
1.1 Zielsetzung und methodisches Vorgehen 1
1.2 Aufbau der Arbeit 2
2 Theoretische Grundlagen 3
2.1 Bestehende Arbeiten 3
2.2 Neuronales TTS und Zero-Shot Voice Cloning 4
2.3 Qualitätsbewertung synthetischer Sprachausgaben 5
2.3.1 Mean Opinion Score (*MOS*) 5
2.3.2 UTMOS 6
2.3.3 Word Error Rate (*WER*) 6
2.4 Statistische Auswertungsverfahren 7
2.4.1 Interrater-Reliabilität (*IRR*) 7
2.4.2 Bayessche Ordinale Logistische Regression 8
2.4.3 Validierung und Vergleich bayesscher Modelle 9
3 Methodik 10
3.1 Annotationsdurchführung 10
3.1.1 Technische Umsetzung 10
3.1.2 Onboarding 11
3.1.3 Kalibrierung und Annotationsreihenfolge 12
3.1.4 Annotation 14
3.2 Auswahl der Referenzstimmen 15
3.3 Auswahl der Sätze 16
3.4 TTS-Pipeline und Qualitätssicherung 17
3.4.1 Generierung der Stimuli 18
3.4.2 Filterung der Stimuli 18
3.4.3 Automatische Bewertung der Stimuli 19
3.5 Modellauswahl 19
3.6 Generierung und Analyse des finalen Datensatzes 22
4 Annotationsstudie 25
4.1 Teilnehmende 25
4.2 Kalibrierung 28
4.2.1 Kalibrierungsresultate 28
4.2.2 Eignung der Kalibrierungsaufnahmen 30
4.2.3 Korrelationsanalyse von Sicherheit und Korrektheit 31
4.2.4 Eingewöhnungseffekt 32
4.3 Annotationen 33

5 Auswertungsmethoden 36
5.1 Majority Vote 36
5.1.1 IRR 37
5.2 Spezifikation der bayesschen Modelle 38
5.2.1 Modellgruppe M1: Personenspezifischer Bias 39
5.2.2 Modellgruppen M2, M3: Stimulusspezifische Effekte 40
5.2.3 Modellgruppe M4: Annotationssicherheit 41
5.2.4 Modellgruppe M5: Heimdialekt 42
5.2.5 Modellgruppe M6: Selbsteinschätzung 42
5.2.6 Modellgruppe M7: Kalibrierungsscore 43
5.2.7 Modellgruppe M8: Altersgruppen 44
5.2.8 Modellgruppe M9: Dialektexposition 45
6 Resultate 47
6.1 Modellverifikation und Prädiktionsgüte 47
6.2 Relative Relevanz der Modellkomponenten 48
6.3 Analyse der stimulusspezifischen Effekte 50
6.3.1 Dialekteffekte 50
6.3.2 Referenzstimmeffekte 52
6.3.3 Satzeffekte 55
6.4 Analyse der personenspezifischen Effekte 58
6.4.1 Relevante Effekte: Annotationssicherheit und Selbsteinschätzung 58
6.4.2 Vernachlässigbare Effekte: Heimdialekt, Kalibrierungsscore, Alter, Dialektexposition 60
7 Diskussion 64
7.1 Reflexion der Ergebnisse 64
7.2 Limitationen 65
7.3 Ausblick 65
8 Verzeichnisse 67
8.1 Literaturverzeichnis 67
8.2 Abbildungsverzeichnis 69
8.3 Tabellenverzeichnis 70
9 Anhang 71
9.1 Einsatz künstlicher Intelligenz (KI) 71
9.2 Quellcode und technische Dokumentation 71

1 Einleitung

Text-To-Speech-Systeme (*TTS*) haben in den letzten Jahren enorme Fortschritte erzielt und erreichen insbesondere im Bereich der Voice Adaptation zunehmend ein menschliches Qualitätsniveau [1]. Parallel dazu ist die Anzahl wissenschaftlicher Publikationen zur Voice Adaptation in den letzten Jahren deutlich gestiegen, was auf ein allgemein wachsendes Forschungsinteresse in diesem Bereich hinweist [2]. Moderne neuronale End-to-End-Architekturen liefern bereits hochqualitative Resultate für standardisierte Hochsprachen. Low-Resource-Sprachen wie Schweizerdeutsch stellen die Sprachsynthese jedoch aufgrund limitierter Datensätze und annotierter Sprachkorpora sowie einer geringen Anzahl Sprechender weiterhin vor komplexe Herausforderungen [3]. Die linguistische Situation in der Schweiz ist durch eine mediale Diglossie geprägt, bei der in der Alltagskommunikation fast ausschliesslich Dialekt gesprochen, als Schriftsprache jedoch Standarddeutsch verwendet wird [4]. Da dem Schweizerdeutschen zudem eine normierte Rechtschreibung fehlt, müssen TTS-Modelle den standarddeutschen Eingabetext nicht nur akustisch synthetisieren, sondern gleichzeitig in den gewünschten Dialekt übersetzen [5].

In vorangegangenen Arbeiten am betreuenden Institut (*Centre for Artificial Intelligence, ZHAW*) wurden für diese Problemstellung Zero-Shot-Voice-Adaptation-Modelle entwickelt, die gezielt auf sieben schweizerdeutsche Dialektregionen verfeinert wurden [3], [6]. Diese Systeme sind in der Lage, auf Basis einer kurzen Referenzaufnahme eines unbekannten Sprechenden (*Referenzstimme*) einen standarddeutschen Text in einen schweizerdeutschen Dialekt der sieben Dialektregionen zu synthetisieren [3]. Obwohl diese Modelle in ersten Evaluationen sehr hohe Werte für die Sprecherähnlichkeit (*SMOS*) und Audioqualität (*CMOS*) erzielten [6], besteht dennoch eine wesentliche methodische Forschungslücke: Die Evaluation der tatsächlichen Dialektauthentizität erfolgte bislang vorwiegend anhand automatisierter, phonembasierter Klassifikatoren [6]. Eine Validierung dieser Ergebnisse durch menschliche Annotationen liegt bislang jedoch nur begrenzt vor. Zudem ist die Qualität der Dialektzugehörigkeit insgesamt noch unzureichend erforscht.

1.1 Zielsetzung und methodisches Vorgehen

Genau an dieser Forschungslücke setzt die vorliegende Bachelorarbeit an. Das primäre Ziel dieser Arbeit ist die Entwicklung und Anwendung eines systematischen Verfahrens zur empirischen Evaluierung der Dialektauthentizität schweizerdeutscher TTS-Systeme. Im Zentrum steht dabei nicht die Optimierung der Sprachsynthese selbst, sondern die Erschaffung einer robusten Evaluationsmethodik, welche die subjektive menschliche Wahrnehmung quantifizierbar macht.

Um dieses Ziel zu erreichen, ist die Aufgabenstellung in drei zentrale Phasen gegliedert:

- **Forschungsdesign und Datengrundlage:** Zunächst wird das Design der empirischen Annotationsstudie konzipiert. Die Datenbasis wird durch die gezielte Auswahl von Referenzstimmen sowie die Definition spezifischer Sätze geschaffen. Auf dieser Grundlage werden die TTS-Ausgaben mittels einer automatisierten Pipeline generiert und einer systematischen Qualitätsprüfung unterzogen.
- **Entwicklung einer Annotationsplattform:** Zur Durchführung der Studie wird eine dedizierte, webbasierte Annotationsplattform entwickelt.

- **Statistische Auswertung:** Die erhobenen Annotationsdaten werden mithilfe geeigneter statistischer und probabilistischer Methoden ausgewertet. Dabei gilt es insbesondere, die inhärent subjektive Natur menschlicher Bewertungen methodisch zu berücksichtigen.

1.2 Aufbau der Arbeit

Die folgende Gliederung soll der Leserschaft einen klaren Überblick über den Aufbau der Arbeit geben. Jedes Kapitel baut auf dem vorherigen auf und orientiert sich dabei am methodischen Ablauf der Studie:

- **Kapitel 2 Theoretische Grundlagen** ordnet die Arbeit in den aktuellen Forschungsstand ein, erläutert die Funktionsweise neuronaler Sprachsynthese und legt die Grundlagen für die Qualitätsbewertung sowie die statistischen Auswertungsverfahren dar.
- **Kapitel 3 Methodik** beschreibt das Design der Annotationsstudie, die technische Pipeline zur Generierung und Qualitätssicherung der TTS-Ausgaben sowie die systematische Erstellung des finalen Datensatzes für die Annotationsstudie.
- **Kapitel 4 Annotationsstudie** analysiert die erhobenen Daten und betrachtet die demografische Struktur sowie die Kalibrierungsresultate der Teilnehmenden.
- **Kapitel 5 Auswertungsmethoden** führt mit der Interrater-Reliabilität und der bayesschen Regression zwei Ansätze zur vertieften Datenanalyse ein.
- **Kapitel 6 Resultate** wendet die Auswertungsmethoden an und quantifiziert die verschiedenen Einflüsse auf das Bewertungsverhalten der Teilnehmenden.
- **Kapitel 7 Diskussion** schliesst die Arbeit mit einer fundierten Reflexion der Ergebnisse ab, beleuchtet methodische Limitationen und bietet einen Ausblick auf zukünftige Forschungspotenziale.

2 Theoretische Grundlagen

Das folgende Kapitel legt das theoretische und methodische Fundament für die empirische Untersuchung dieser Arbeit. Dazu wird zunächst der aktuelle Forschungsstand beleuchtet und abgegrenzt. Darauf aufbauend werden die Mechanismen der neuronalen Sprachsynthese sowie des Zero-Shot Voice Cloning dargelegt, welche die Basis für die Generierung der TTS-Ausgaben bilden. Ein weiterer Schwerpunkt liegt auf der systematischen Qualitätsbewertung synthetischer Sprachausgaben mittels subjektiver und automatisierter Metriken. Das Kapitel schliesst mit der theoretischen Herleitung der statistischen Auswertungsverfahren ab, die der späteren Modellierung, Bereinigung und Validierung der erhobenen Annotationsdaten dienen.

2.1 Bestehende Arbeiten

Die Erforschung und Verarbeitung von Schweizerdeutsch im Bereich des Natural Language Processing (*NLP*) hat in den letzten Jahren stark an Bedeutung gewonnen. Da es sich um eine primär gesprochene Sprache ohne standardisierte Schriftsprache und mit ausgeprägter dialektaler Vielfalt handelt, stellt Schweizerdeutsch eine klassische Low-Resource-Herausforderung dar. Der Fokus der bisherigen Forschung liegt auf der Schaffung von Datengrundlagen, der automatischen Dialekterkennung sowie jüngsten Fortschritten in der Sprachsynthese.

Ressourcen und Datensätze
Um datenintensive Modelle trainieren zu können, ist die Verfügbarkeit hochwertiger Korpora unerlässlich. Einen massgeblichen Beitrag lieferte hierbei die Publikation des SwissGPC v1.0 Corpus. Ziel dieser Arbeit war die Schaffung des ersten mittel- bis grossskaligen Korpus für spontanes Schweizerdeutsch, um die Forschung in den Bereichen der automatischen Spracherkennung (*ASR*) und der Sprachsynthese (*TTS*) zu unterstützen. Über eine automatisierte Pipeline wurden dafür mehr als 5400 Stunden Rohaudio gesammelt, transkribiert, segmentiert und in sieben Dialektregionen klassifiziert. Obwohl das Resultat ein umfassender Datensatz unkontrollierter Konversationssprache ist, bleibt eine wesentliche Limitation die starke Ungleichverteilung der Daten: Zürichdeutsch und Standarddeutsch dominieren das Korpus mit fast 60 Prozent, während Dialekte wie Walliserdeutsch mit unter einem Prozent stark unterrepräsentiert sind. [7]

Automatische Dialekterkennung und Sprachsynthese
Während im Bereich der automatischen Dialekterkennung und Sprachverarbeitung (*z. B. mittels Whisper oder Wav2Vec2*) bereits signifikante Fortschritte bei der Klassifizierung schweizerdeutscher Audiodaten erzielt wurden, verlagert sich der aktuelle Forschungsfokus zunehmend auf die generative Sprachsynthese und Voice Adaptation. [8], [9], [10], [11]

Aufbauend auf diesen Grundlagen konzentrierten sich jüngste Arbeiten auf die Entwicklung von Zero-Shot-TTS-Modellen für Schweizerdeutsch. Dabei wurde das vortrainierte, multilinguale Modell XTTSv2 mittels Fine-Tuning an Schweizerdeutsch angepasst. Das System nimmt standarddeutschen Text als Eingabe und generiert daraus hochqualitative Sprachausgaben im gewünschten Dialekt mit der Stimme eines unbekannten Referenzsprechenden. Die Evaluierung zeigte vielversprechende Ergebnisse und belegte die generelle Machbarkeit von Zero-Shot-TTS für Low-Resource-Dialekte mit niedriger Word Error Rate (*WER*). Eine methodische Limitation dieser Synthesearbeiten liegt jedoch in der menschlichen Evaluierung: Diese war meist auf sehr kleine, wenig repräsentative Nutzergruppen beschränkt und beruhte auf deskriptiver oder simpler vergleichender Statistik. Zudem wird in der Literatur selbst angemerkt, dass

menschliche Evaluierungen der Dialektgenauigkeit oft unzuverlässig sind und stark von regionalen Verzerrungen der Annotierenden abhängen. [3], [6]

Die Literatur zeigt somit fundamentale Fortschritte bei der automatischen Verarbeitung von Schweizerdeutsch. Trotz dieser technologischen Durchbrüche auf der Syntheseseite fehlt bislang eine systematische und methodisch tiefgreifende Evaluierung der tatsächlichen Dialektauthentizität. Bisherige Arbeiten evaluierten ihre Pipelines vorwiegend anhand automatisierter Metriken oder einfacher Mean Opinion Scores (*MOS*) in kleinen Gruppen, ohne komplexe subjektive Verzerrungen statistisch abzufangen.

2.2 Neuronales TTS und Zero-Shot Voice Cloning

Dieses Kapitel vermittelt das technologische Fundament der in dieser Arbeiten genutzten TTS-Pipeline. Zunächst wird aufgezeigt, wie sich die Sprachsynthese zu neuronalen End-to-End-Systemen entwickelt hat. Anschliessend folgen eine Beschreibung der Architektur des XTTS-Frameworks sowie eine Erörterung der systembedingten Limitationen des Zero-Shot-Ansatzes.

Entwicklung zu neuronalen End-to-End-Modellen: In der Vergangenheit basierten TTS-Systeme primär auf der konkatenativen Synthese, bei der ein Suchalgorithmus vorab aufgenommene Sprachsegmente dynamisch zusammensetzte. Dieser Ansatz war jedoch speicherintensiv, unflexibel und führte zu akustischen Diskontinuitäten. Moderne neuronale End-to-End-Architekturen lösen dieses Problem, indem sie das nichtlineare Mapping zwischen Texteingabe und akustischer Waveform direkt innerhalb eines Netzwerks erlernen. Der Prozess ist meist zweistufig: Ein Feature-Generator übersetzt Text-Tokens in ein Mel-Spektrogramm, das ein neuronaler Vocoder anschliessend in ein kontinuierliches Audiosignal umwandelt. [12]

Zero-Shot Voice Cloning: Im Gegensatz zu klassischen Modellen, die ein aufwendiges Fine-Tuning für jede neue Referenzstimme erfordern, erlaubt der Zero-Shot-Ansatz die Imitation einer unbekannten Stimme, ohne erneutes Training des Basismodells [13]. Wie in Abbildung 1 schematisch dargestellt, extrahiert ein Encoder aus einer kurzen Referenzaufnahme das charakteristische Profil der Zielstimme (*Speaker Embedding*). Dieses Profil dient im Synthesizer als zusätzliche Bedingung für den Text, sodass das Modell das finale Audiosignal mit genau diesen stimmlichen Eigenschaften generiert [13]. Obwohl für diesen Prozess theoretisch wenige Sekunden Audio ausreichen, erzielen im Dialektbereich längere Referenzen (*ca. 25 Sekunden*) eine deutlich höhere Stimmähnlichkeit [3].

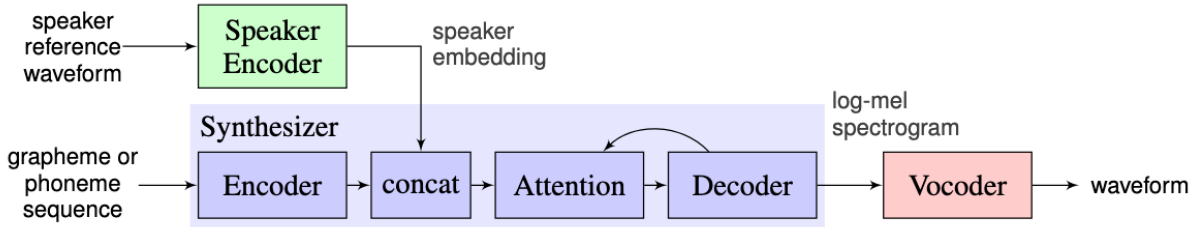

Abbildung 1: Schematische Darstellung einer Zero-Shot-TTS-Architektur mit separatem Speaker Encoder zur Extraktion von Stimmmerkmalen. Der Datenfluss verläuft von lins nach rechts durch die drei Kernkomponenten: Speaker Encoder (oben), Synthesizer (unten links) und Vocoder (unten rechts). Entnommen aus [13]

Die XTTS-Architekturkomponenten: Das in dieser Arbeit genutzte Framework XTTSv2 (*Coqui*) verwendet eine sprachmodellbasierte Architektur [14]. Das System lässt sich auf drei Kernschritte reduzieren:

1. **Sprachmodellierung:** Das Modell übersetzt Audio in diskrete Tokens. Ein integriertes Sprachmodell (*basierend auf der GPT-Architektur*) berechnet anschliessend aus dem eingegebenen Text die passenden Audio-Tokens [3], [14].
2. **Conditioning (*Stimmextraktion*):** Ein spezialisierter Encoder extrahiert das Speaker Embedding aus der Referenzaufnahme. Ein Optimierungsverfahren (*Speaker Consistency Loss*) sorgt während der Generierung dafür, dass die Stimmidentität über den gesamten Satz hinweg stabil bleibt [3].
3. **Vocoder (*Audio-Generierung*):** Abschliessend wandelt ein Vocoder die berechneten Tokens in ein kontinuierliches Audiosignal mit einer Abtastrate von 24 kHz um, was eine hohe klangliche Detailtreue der synthetisierten Stimme ermöglicht [3], [14].

Technologische Limitationen: Die Anwendung von Zero-Shot-TTS-Modellen bringt drei wesentliche systembedingte Limitationen mit sich:

- **Qualität der Referenzaufnahme:** Speaker Encoder übernimmt Störgeräusche wie Hintergrundrauschen oder Raumhall direkt in das generierte Audio [3].
- **Satzlänge:** Bei sehr kurzen Sätzen (*unter 5 bis 7 Sekunden*) neigt das XTTS-Modell aufgrund der Struktur seiner Trainingsdaten zu Instabilitäten. Dies kann zu Artefakten oder abgehacktem Audio führen [3].
- **Dialektvielfalt:** Bei stark abweichenden oder unterrepräsentierten Dialekten (*Randdialekte*) sinkt erwartungsgemäss die Natürlichkeit der Synthese, während die Fehlerquote steigt [3].

2.3 Qualitätsbewertung synthetischer Sprachausgaben

Zur Evaluation von TTS-Systemen werden in der Forschung sowohl subjektive als auch automatisierte Metriken eingesetzt. Da die menschliche Wahrnehmung von Nuancen, wie der Natürlichkeit oder der Dialektauthentizität, hochkomplex ist, gelten strukturierte Hörexperimente mit Probanden weiterhin als der Goldstandard [15]. Weil die Durchführung solcher Annotationsstudien jedoch zeit- und kostenintensiv ist, kommen ergänzend automatisierte Verfahren zum Einsatz, um die Qualität der generierten Audiodaten objektiv, skalierbar und reproduzierbar zu messen.

Die folgenden Abschnitte stellen die drei zentralen Bewertungsmetriken dieser Arbeit vor. Zunächst wird das Konzept des MOS als etablierter Standard für die subjektive Bewertung erläutert. Anschliessend werden zwei automatisierte Metriken behandelt: die WER zur Messung der akustischen Verständlichkeit sowie der UTMOS, ein neuronales Modell zur maschinellen Vorhersage der Sprachqualität.

2.3.1 Mean Opinion Score (*MOS*)

Die am weitesten verbreitete Methode zur Quantifizierung subjektiver Höreindrücke ist der MOS. Dieses ursprünglich für die Telekommunikation entwickelte Verfahren ist durch die International Telecommunication Union im Standard ITU-T Recommendation P.800 formalisiert [16]. Bei der standardisierten Durchführung mittels Absolute Category Rating (*ACR*) bewerten die Versuchspersonen isolierte Audioaufnahmen auf einer diskreten Skala. Traditionell reicht

diese fünfstufige Skala von 1 (*schlecht*) bis 5 (*ausgezeichnet*). Der finale MOS-Wert eines Systems berechnet sich aus dem arithmetischen Mittelwert aller Einzelbewertungen.

Je nach Studienziel unterscheidet man verschiedene akustische Dimensionen:

- **Audioqualität:** Beurteilung der reinen Signalintegrität (*z. B. Abwesenheit von Rauschen, Verzerrungen oder digitalen Artefakten*).
- **Natürlichkeit:** Beurteilung, wie menschlich und organisch die generierte Stimme unabhängig vom Inhalt klingt.
- **Authentizität:** In spezialisierten Anwendungsfällen, wie in der vorliegenden Arbeit, wird das Konzept adaptiert, um die Glaubwürdigkeit eines spezifischen Dialekts zu evaluieren.

Da subjektive Umfragen anfällig für individuelle Verzerrungen (*Bias*) oder unaufmerksames Antwortverhalten sind, ist eine strenge Qualitätskontrolle erforderlich. Eine gängige Praxis ist es, neben der Zielmetrik (*wie der Authentizität*) auch die subjektive Sicherheit der Annotierenden abzufragen.

2.3.2 UTMOS

Der UTokyo-SaruLab Mean Opinion Score (*UTMOS*) wird in dieser Arbeit als primäre Metrik für die automatische Qualitätsbewertung der TTS-Ausgaben verwendet. UTMOS basiert auf einem neuronalen Netzwerk, das darauf trainiert wurde, den menschlichen MOS prädiktiv zu schätzen. Die Metrik liefert kontinuierliche Werte auf einer Skala von 1 bis 5 und bietet eine reproduzierbare Alternative zum traditionellen Hörexperiment. Bei der VoiceMOS Challenge 2022 erzielte UTMOS die höchsten Korrelationswerten mit menschlichen Urteilen und gilt seither als zuverlässiges System zur automatischen Evaluierung von TTS-Ausgaben. [17]

2.3.3 Word Error Rate (*WER*)

Die WER dient in dieser Arbeit als sekundäre Metrik für die automatische Qualitätsbewertung der TTS-Ausgaben. Während diese Metrik klassischerweise primär zur Evaluation von automatischen Spracherkennungssystemen (*ASR/STT*) eingesetzt wird, erfüllt sie hier den Zweck, die inhaltliche Genauigkeit der Sprachsynthese zu verifizieren. Durch die automatische Transkription der generierten TTS-Ausgaben lässt sich objektiv messen, ob das TTS-Modell den vorgegebenen Text tatsächlich korrekt und verständlich ausgesprochen hat. Die WER misst die Abweichung zwischen dem ursprünglichen Referenztext und der automatischen Transkription einer Audioausgabe auf Wortebene. Die Berechnung erfolgt nach der folgenden Gleichung:

$$\mathrm{WER} = \frac{S + D + I}{N}$$

Hierbei steht S für die Anzahl substituierten Wörter (*Substitutions*), D für ausgelassene Wörter (*Deletions*), I für fälschlicherweise hinzugefügte Wörter (*Insertions*) und N für die Gesamtanzahl der Wörter des Referenztextes. Da die Anzahl der Insertions nicht durch die Länge des Referenztextes beschränkt ist, kann der resultierende Wert theoretisch unbegrenzt wachsen und somit auch Werte über 1 annehmen.

2.4 Statistische Auswertungsverfahren

Während automatische Metriken eine effiziente Basisbewertung ermöglichen, erfordert die Beurteilung subjektiver Merkmale, wie der Dialektauthentizität, menschliche Einschätzungen. Die statistische Auswertung dieser Annotationen bringt jedoch spezifische Herausforderungen mit sich. Dieses Kapitel behandelt zunächst die Messung der Übereinstimmung zwischen den Teilnehmenden, bevor die Modellierung der ordinalen Bewertungsdaten mittels bayesscher Regression erläutert wird.

2.4.1 Interrater-Reliabilität (*IRR*)

Bei Annotationsstudien mit mehreren Teilnehmenden ist die Quantifizierung der Datenkonsistenz essenziell. Ein einfacher Mehrheitsentscheid (*Majority Vote*), bei dem der am häufigsten genannte Wert als finales Resultat dient, stösst bei subjektiven Merkmalen an Grenzen: Variiert die Wahrnehmung stark, verdeckt dieser Ansatz legitime Diskrepanz und suggeriert eine unberechtigte Eindeutigkeit. Zur robusten Quantifizierung der Übereinstimmung wird daher die Interrater-Reliabilität (*IRR*) bestimmt.

Ein etabliertes Mass für die IRR ist Krippendorffs Alpha (α). Diese Metrik ist robust gegenüber unterschiedlichen Skalenniveaus, fehlenden Werten und einer variablen Anzahl an Annotierenden pro Item. Die Grundgleichung setzt die beobachtete Disagreement D_o ins Verhältnis zum erwarteten Disagreement D_e bei zufälliger Verteilung:

$$\alpha = 1 - \frac{D_o}{D_e}$$

Nach Krippendorff gelten Daten mit $\alpha \geq 0.8$ als zuverlässig, während Werte im Bereich von $0.67 \leq \alpha < 0.8$ für vorläufige Schlussfolgerungen akzeptabel sind [18]. Werte unter 0.67 weisen auf eine unzureichende Übereinstimmung hin, bei der die Teilnehmenden den Items keine konsistenten Labels zuweisen konnten [18]. Die Berechnung der Disagreements D_o und D_e erfolgt über die Gleichungen:

$$D_o = \frac{1}{\sum_i m_i(m_i - 1)} \sum_i \sum_{c \neq k} n_{ick} \cdot \delta^2(c,k)$$

$$D_e = \frac{1}{N(N-1)} \sum_{c \neq k} n_c \cdot n_k \cdot \delta^2(c,k)$$

Hierbei repräsentiert m_i die Anzahl der Bewertungen für die Audioausgabe i und n_{ick} für die Anzahl der Wertepaare (c,k) innerhalb dieser Audioausgabe i. Die Gesamthäufigkeit der Werte c und k über alle Bewertungen hinweg wird durch n_c respektive n_k definiert, während N die Gesamtzahl aller Bewertungen abbildet. Die Indizes c und k stehen für verschiedene Bewertungsstufen.

Die Bewertung erfordert zudem die Definition einer Distanzfunktion $\delta^2(c,k)$, welche den Abstand zwischen den Bewertungsstufen interpretiert. Während eine ordinale Distanz nur die Rangfolge berücksichtigt, setzt die in der Praxis oft genutzte Intervalldistanz eine Äquidistanz voraus. Sie nimmt somit an, dass die Abstände zwischen allen benachbarten Antwortkategorien exakt identisch sind [19]. Unter dieser Annahme berechnet sich die quadrierte normalisierte Intervalldistanz wie folgt:

$$\delta^2(c,k) = \left(\frac{|c-k|}{\max(r) - \min(r)}\right)^2$$

Diese Distanzmetrik bestraft grosse Abweichungen zwischen den Annotierenden und bildet die Basis der Annotationskonsistenz.

2.4.2 Bayessche Ordinale Logistische Regression

Im Gegensatz zu direkten Aggregationsmethoden wie dem Majority Vote operiert die bayessche ordinale logistische Regression nicht direkt auf den beobachteten Bewertungen. Stattdessen beschreibt ein probabilistisches Modell die zugrunde liegenden Einflussfaktoren auf das Antwortverhalten mathematisch.

Da die Zielvariable ordinal ist, wird eine latente kontinuierliche Variable y^* angenommen, welche die tatsächlich wahrgenommene Qualität einer Audioausgabe repräsentiert. Diese latente Variable wird als lineare Kombination der Einflussfaktoren modelliert:

$$y_i^* = \mu_i + \epsilon_i$$

Dabei beschreibt μ_i den deterministischen Mittelwert (*linearen Prädikator*), der sich aus den Modelleingangsparameter zusammensetzt, während ϵ_i den Fehlerterm abbildet. Schwellenwerte τ (*Cutpoints*) unterteilen diese latente Variable in die diskreten, beobachtbaren Kategorien y_i:

$$y_i = k \quad \text{falls} \quad \tau_{k-i} < y_i^* \leq \tau_k$$

Die Parameter τ_k definieren somit die Übergänge zwischen den Bewertungsklassen. Unter Verwendung einer logistischen Link-Funktion ergeben sich die kumulativen Wahrscheinlichkeiten als:

$$P(y_i \leq k) = \frac{1}{1 + e^{-(\tau_k - \mu_i)}}$$

Die Wahrscheinlichkeit für eine diskrete Kategorie ergibt sich anschliessend aus der Differenz benachbarter kumulativer Wahrscheinlichkeiten.

Im bayesschen Paradigma werden alle unbekannten Parameter nicht als Punktwerte, sondern als Zufallsvariablen mit eigenen Wahrscheinlichkeitsverteilungen modelliert. Dem Prädikator β liegt meist eine Normalverteilung zugrunde, während die Schwellenwerte τ strikt monoton geordnet sein müssen ($\tau_1 < \tau_2 < \ldots < \tau_K$). Die Inferenz basiert auf der Posterior-Verteilung, die sich proportional aus Likelihood und Priors zusammensetzt:

$$p(\beta, \tau \mid y) \propto p(y \mid \beta, \tau)\, p(\beta)\, p(\tau)$$

Diese Posterior-Verteilung wird mittels Markov Chain Monte Carlo (*MCMC*) approximiert. Dieses stochastische Verfahren erzeugt eine Markov-Kette, deren stationäre Verteilungen der gesuchten Posterior-Verteilungen entspricht. Durch wiederholtes Ziehen von Stichproben wird die Verteilung numerisch angenähert, was eine vollständige Quantifizierung der Parameterunsicherheit ermöglicht. [20]

2.4.3 Validierung und Vergleich bayesscher Modelle

Die Evaluation von MCMC-approximierten Modellen unterscheidet sich grundlegend von frequentistischen Ansätzen. Der Validierungsprozess gliedert sich in drei Schritte: die Überprüfung der Sampling-Qualität (*Konvergenzdiagnostik*), die Kontrolle der strukturellen Modelladäquatheit sowie den quantitativen Modellvergleich.

Konvergenzdiagnostik: Da die Posterior-Verteilung stochastisch approximiert wird, muss die Konvergenz der MCMC-Ketten sichergestellt sein. Ein zentrales Gütekriterium ist der Potential Scale Reduction Factor *($\hat{R}$, auch RHAT genannt)*. Dieser Faktor vergleicht die Varianz innerhalb einzelner Sampling-Ketten mit der Varianz zwischen unabhängigen Ketten. Konvergieren die Ketten gegen dieselbe Zielverteilung, nähert sich $\hat{R}$ dem Wert 1.0 an. In der Praxis gilt ein Schwellenwerte von $\hat{R} \leq 1.01$ als Beleg für eine erfolgreiche Konvergenz. [21]

Ergänzend wird die Effective Sample Size (*ESS*) herangezogen. Aufgrund der inhärenten Autokorrelation von MCMC-Verfahren entspricht die reine Stichprobenanzahl nicht der Menge unabhängiger Informationen. Die ESS quantifiziert diese effektive Informationsmenge, wobei Werte ab 400 als ausreichend für verlässliche Inferenzschätzungen gelten. [21]

Posterior Predictive Checks (*PPC*): Um zu prüfen, ob das trainierte Modell die zugrunde liegende Datenstruktur korrekt erfasst, werden PPCs durchgeführt. Dabei werden aus den berechneten Posterior-Verteilungen künstliche Datensätze simuliert und mit den empirischen ordinalen Klassenhäufigkeiten abgeglichen. Systematische Abweichungen zwischen der simulierten und der beobachteten Verteilungen deuten auf eine Fehlspezifikation des Modells hin, beispielsweise durch unberücksichtigte Einflussfaktoren oder ein unzureichend modelliertes Klassenungleichgewichten.

Informationskriterien und Modellvergleich: Der quantitative Vergleich konkurrierender Modelle erfolgt über deren prädiktive Genauigkeit mittels der Leave-One-Out Cross-Validation (*LOO-CV*). Diese schätzt die erwartete logarithmische punktweise prädiktive Dichte (*Expected Log Pointwise Predictive Density, ELPD*) für ungesehene Datenpunkte. Höhere ELPD-Werte signalisieren eine bessere prädiktive Leistung.

Die LOO-CV bestraft Overfitting automatisch und quantifiziert so den tatsächlichen informationellen Mehrwert zusätzlicher Prädiktoren. Zudem wird die Stabilität der geschätzten Effekte (*Posterior Effect Stability*) bei der Hinzunahme weiterer Variablen untersucht: Bleiben die Effekte stabil, deutet dies auf Unabhängigkeit hin; starke Schwankungen weisen hingegen auf eine Korrelation der Prädiktoren (*Confounding*) hin.

3 Methodik

Dieses Kapitel beschreibt die methodischen Grundlagen für die Erstellung und Evaluation synthetischer schweizerdeutscher Sprachausgaben. Der Aufbau orientiert sich dabei an den zentralen Bausteinen der Studie:

Zunächst wird das Forschungsdesign der empirischen Annotationsstudie dargelegt, welches die technische Umsetzung sowie das Onboarding und die Kalibrierung der Teilnehmenden umfasst. Anschliessend wird die zugrunde liegende Datenbasis definiert, bestehend aus der Selektion geeigneter Referenzstimmen und der Festlegung dialektspezifischer Sätze. Darauf aufbauend wird die automatisierte Software-Pipeline zur Generierung und Qualitätssicherung der TTS-Ausgaben vorgestellt. Sie bildet das technische Fundament für das darauffolgende systematische Verfahren zur Modellauswahl. Den Abschluss des Kapitels bildet die Generierung und statistische Vorverarbeitung des finalen Datensatzes, der als Grundlage für die spätere Auswertung dient.

3.1 Annotationsdurchführung

Dieser Abschnitt beschreibt die technische Umsetzung und methodische Gestaltung der Annotationsstudie. Die Implementierung erfolgte als webbasierte Plattform und gliedert sich in drei Phasen:

1. **Onboarding:** Erhebung demografischer und dialektbezogener Informationen der Teilnehmenden.
2. **Kalibrierung:** Einschätzung der Dialekterkennungskompetenz der Teilnehmenden anhand vorannotierter Referenzaufnahmen.
3. **Annotation:** Finale Bewertung der generierten TTS-Ausgaben durch die Teilnehmenden.

3.1.1 Technische Umsetzung

Die Annotationsplattform wurde im Rahmen dieser Bachelorarbeit als eigenständige Webanwendung implementiert. Der Technologie-Stack umfasst das Webframework Next.js mit TypeScript, die dateibasierte Datenbank SQLite sowie das Framework Drizzle für das objektrelationale Mapping (*ORM*).

Die Teilnehmenden können sich selbstständig mittels E-Mail-Adresse und Passwort registrieren. Ein persönliches Benutzerkonto stellt sicher, dass Antworten und Annotationen fortlaufend zwischengespeichert werden, was eine flexible Unterbrechung und spätere Fortsetzung der Studie erlaubt.

Ergänzend verfügt die Plattform über ein Administrations-Backend zur Verwaltung von Studien, Benutzerdaten und Resultaten. Dadurch lassen sich neue Erhebungen konfigurieren und die erhobenen Daten exportieren, ohne dass programmiertechnische Anpassungen am Quellcode nötig sind. Der vollständige Quellcode der Plattform ist in Anhang 9.1 abrufbar.

3.1.2 Onboarding

Im Onboarding werden vor der eigentlichen Annotation demografische und dialektbezogene Informationen der Teilnehmenden erhoben. Diese dienen dazu, die Annotationsresultate zu kontextualisieren und den Einfluss individueller Eigenschaften auf die Bewertungen zu untersuchen. Das Onboarding umfasst vier Teilschritte:

1. **Altersgruppe**: Die Teilnehmenden ordnen sich über ein Dropdown-Menü einer der sieben verfügbaren Alterskategorien zu (*Unter 18 Jahre, 18–25 Jahre, 26–35 Jahre, 36–45 Jahre, 46–55 Jahre, 56–65 Jahre, Über 65 Jahre*).
2. **Aufenthaltsorte:** Die Teilnehmenden geben an, in welchen Kantonen der Deutschschweiz sie länger als ein Jahr gelebt haben (*inklusive Aufenthaltsdauer*). Kantone ausserhalb der Deutschschweiz sowie Auslandsaufenthalte können gesammelt unter «Anderer Kanton/Ausland» erfasst werden. Es sind mehrere Einträge möglich.
3. **Dialektregionen:** Die Teilnehmenden wählen aus den sieben Dialektregionen jene aus, der sie sich am meisten zugehörig fühlen. Als Orientierungshilfe dient eine Karte, welche die Kantone den jeweiligen Regionen zuordnet (*siehe Abbildung 2*). Da der Kanton Aargau keiner Dialektregion eindeutig zugeordnet werden kann, werden Betroffene aufgefordert, die passendste benachbarte Region zu wählen[1].
4. **Selbsteinschätzung:** Die Teilnehmenden bewerten pro Dialektregion ihre Fähigkeit, den jeweiligen Dialekt zu erkennen. Dies erfolgt auf einer vierstufigen Skala (*1 = Sehr schlecht, 2 = Schlecht, 3 = Gut, 4 = Sehr gut*). Für unbekannte Dialekte steht die Option „Nicht vertraut" zur Verfügung (*siehe Abbildung 3*).

Abbildung 2: Karte der schweizerdeutschen Dialektregionen mit Kantonszuweisungen. Farbige Flächen zeigen die sieben klassifizierten Dialektregionen, Kürzel stehen für die Kantone. Der Aargau (AG) ist als schraffierte Übergangszone dargestellt. Graue Flächen markieren Gebiete ausserhalb des untersuchten Sprachraums.

[1] In der Forschungspraxis ist die fehlende eindeutige Dialektzuordnung des Kantons Aargau bekannt. Bei der Aufbereitung des SDS-200-Korpus wurden Aufnahmen aus Kantonen mit mehreren möglichen Subregionszuordnungen (darunter Aargau) verworfen [9]. In verwandten Arbeiten wurde der Aargau pragmatisch mit der Region Zürich zusammengefasst [8].

Abbildung 3: Selbsteinschätzungs-Matrix (Dialektidentifikation). Die Teilnehmenden bewerten ihre Erkennungskompetenz pro Dialektregion (Zeilen) auf einer Skala von 1 bis 4 (Spalten). Die Option «N» dient als Ausweichmöglichkeit («keine Angabe»).

3.1.3 Kalibrierung und Annotationsreihenfolge

In der Kalibrierungsphase wird die individuelle Kompetenz der Teilnehmenden zur Dialekterkennung quantitativ eingeschätzt. Die algorithmische Grundannahme dabei lautet: Bewertungen von Personen, die einen bestimmten Dialekt zuverlässig erkennen, haben für dessen spätere Evaluation eine höhere Aussagekraft. Diesen Personen sollen die jeweiligen TTS-Ausgaben in der Annotationsphase früher und häufiger präsentiert werden, um die Gesamtzahl der benötigten Annotationen zu minimieren.

Erstellung des Kalibrierungsdatensatzes
Um diese Kompetenz der Teilnehmenden objektiv zu messen, wurde ein Kalibrierungsdatensatz aus echten Referenzaufnahmen zusammengestellt. Dieser entstammt dem Test-Datensatz des betreuenden Instituts (*Centre for Artificial Intelligence, ZHAW*) mit insgesamt 24’605 Aufnahmen. Die Filterung der Daten erfolgte in drei Stufen:

1. **Dialektabdeckung:** Es wurden ausschliesslich Sätze ausgewählt, die in allen sieben untersuchten Dialektregionen (*Bern, Basel, Graubünden, Innerschweiz, Ostschweiz, Wallis und Zürich*) vorlagen.
2. **Länge:** Um eine ausreichende akustische Beurteilungssbasis zu gewährleisten, wurden alle Audiodateien mit einer Dauer von unter 8 Sekunden verworfen.
3. **Qualität:** Die verbleibenden Dateien durchliefen eine manuelle Kontrolle, bei der Aufnahmen mit Störfaktoren (*Hintergrundgeräusche, unklare Aussprache, Audioartefakte*) aussortiert wurden.

Der resultierende Kalibrierungsdatensatz besteht aus vier Sätzen, was bei sieben Dialektregionen zu genau 28 Aufnahmen führt. Diese werden allen Teilnehmenden in einer festgelegten Reihenfolge präsentiert.

Ablauf der Kalibrierung
Während der eigentlichen Kalibrierungsphase werden pro Aufnahme zwei Parameter erhoben: die vermutete Dialektregion sowie die subjektive Sicherheit dieser Einschätzung (*auf einer vierstufigen Skala*). Um vorschnelles Raten zu unterbinden und die Datenqualität zu sichern, wird die Eingabemaske erst freigeschaltet, nachdem die Referenzaufnahme vollständig abgespielt wurde (*siehe Abbildung 4*).

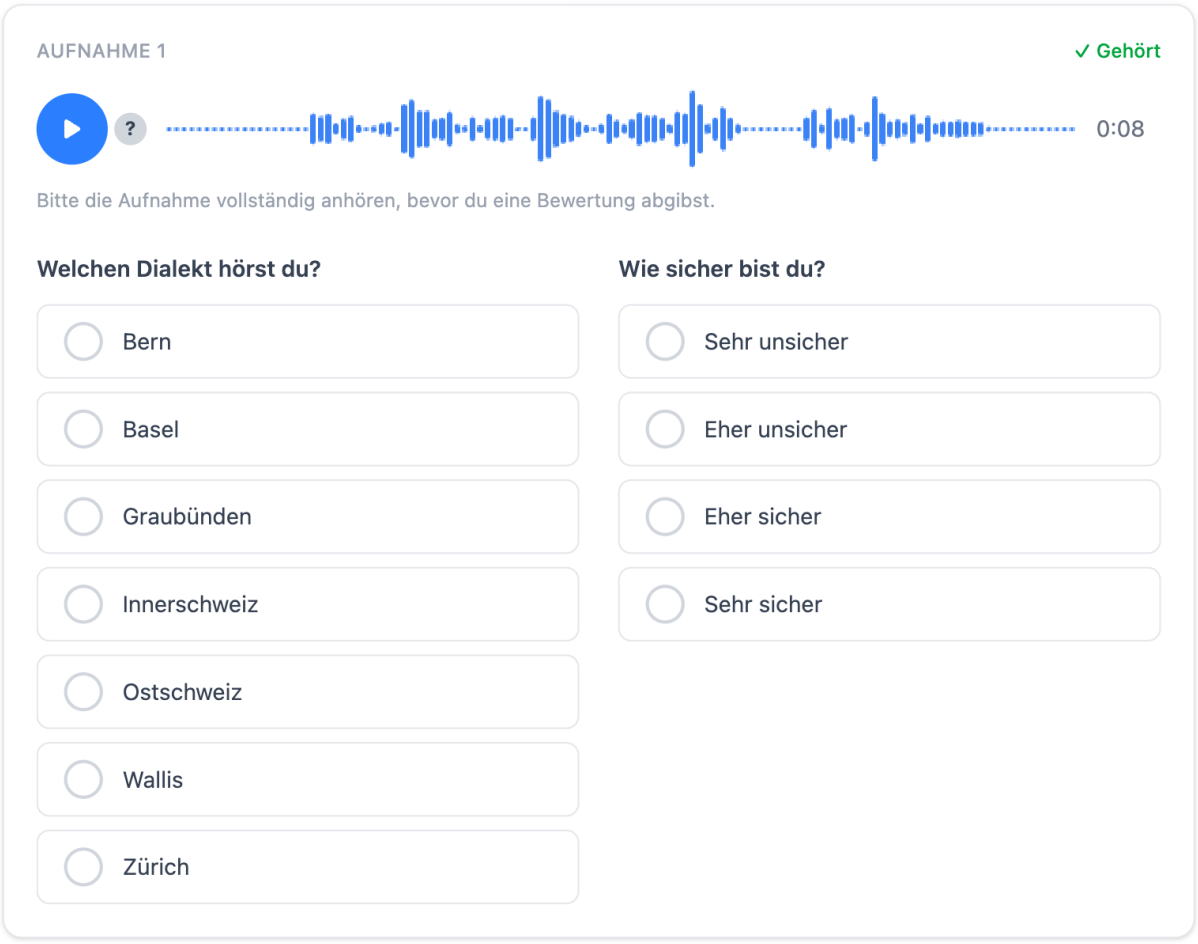

Abbildung 4: Kalibrierungsansicht mit Audioplayer, Dialektauswahl und Sicherheitsskala. Oben befindet sich der Audioplayer zur Wiedergabe des Stimulus. Unten erfolgen zwei Eingaben: Links wird der wahrgenommene Dialekt aus sieben Regionen ausgewählt, rechts die Bewertungssicherheit auf einer vierstufigen Skala erfasst.

Berechnung des Kalibrierungsscores
Aus den 28 Antworten wird pro Dialektregion ein individueller Score berechnet. Zunächst wird die Korrektheit $k \in \{-1,1\}$ definiert (*1 bei korrekter, -1 bei falscher Zuordnung*). Die angegebene Sicherheit $x \in \{1,2,3,4\}$ wird durch Division auf den Bereich $[0,1]$ normalisiert. Der resultierende Score c für eine einzelne Aufnahme berechnet sich somit als:

$$c = \mathrm{k} \cdot \frac{\mathrm{x}}{4}$$

Dieser Wert liegt im Intervall $[-1,1]$, wobei eine korrekte Antwort mit maximaler Sicherheit den Wert 1 und eine falsche Antwort mit maximaler Sicherheit den Wert -1 ergibt. Der aggregierte Dialektscore C_d entspricht anschliessend dem arithmetischen Mittel aller Einzelscores c der jeweiligen Dialektregion.

Gewichtete Annotationszuteilung (*Round-Robin*)
Die berechneten Scores C_d steuern die nachfolgende Zuteilung der Annotationsaufgaben über einen gewichteten Round-Robin-Mechanismus. Jede Dialektgruppe d enthält einen ganzzahligen Gewichtsfaktor $w_d \in \{1, \ldots, 8\}$. Unter Einbezug des höchsten positiven Dialektscores aller Regionen ($C_{\max}$) berechnet sich dieses Gewicht durch kaufmännisches Runden wie folgt:

$$w_d = \max\left(1, \left\lfloor \frac{C_d}{C_{\max}} \cdot 8 \right\rceil\right)$$

In den iterativen Zuteilungsrunden fungiert dieses Gewicht als Multiplikator: Eine Dialektgruppe mit dem Maximalgewicht 8 speist pro Runde achtmal so viele TTS-Ausgaben in den Pool ein wie eine Gruppe mit dem Minimalgewicht 1. Dialekte mit einem negativen Score ($C_d < 0$) sowie Sonderfälle ohne Kalibrierungsscores erhalten standardmässig das Gewicht 1. Dies stellt sicher, dass schwächer erkannte Dialekte zwar seltener präsentiert, aber niemals vollständig herausgefiltert werden.

Um systematische Verzerrungen bei der Auswertung zu vermeiden, wird die finale Reihenfolge deterministisch pro User gemischt. Dadurch erhält jede teilnehmende Person eine reproduzierbare, aber individuell zufällige Abfolge der auszuwertenden TTS-Ausgaben.

3.1.4 Annotation

Der Abschluss des Versuchsprozesses bildet die eigentliche Annotationsphase, in welcher die Teilnehmenden die zu evaluierenden TTS-Ausgaben beurteilen. Diese präsentierten Audiosynthesen werden im weiteren methodischen Verlauf der Arbeit konsequent als Stimuli bezeichnet. Für jeden Stimulus werden zwei unabhängige Dimensionen erhoben (*siehe Abbildung 5*):

1. **Authentizität**: Die Teilnehmenden beurteilen, wie überzeugend der Stimulus nach dem geforderten Dialekt klingt. In Anlehnung an klassische MOS-Erhebungen (siehe Kapitel 2.3.1) erfolgt die Bewertung hier jedoch auf einer vierstufigen Skala (*von 1 = Klingt nicht danach bis 4 = Klingt eindeutig danach*).
2. **Sicherheit:** Zusätzlich geben die Teilnehmenden an, wie sicher sie sich bei ihrem Urteil zur Authentizität sind. Auch dies wird auf einer vierstufigen Skala abgebildet (*von 1 = Sehr unsicher bis 4 = Sehr sicher*).

Zur Sicherstellung der Datenqualität muss die Audioausgabe vollständig angehört werden, bevor eine Bewertung möglich ist. Zur inhaltlichen Unterstützung wird den Teilnehmenden parallel die Texttranskription eingeblendet. Bereits abgegebene Annotationen können im weiteren Verlauf der Studie jederzeit nochmals aufgerufen und bearbeitet werden.

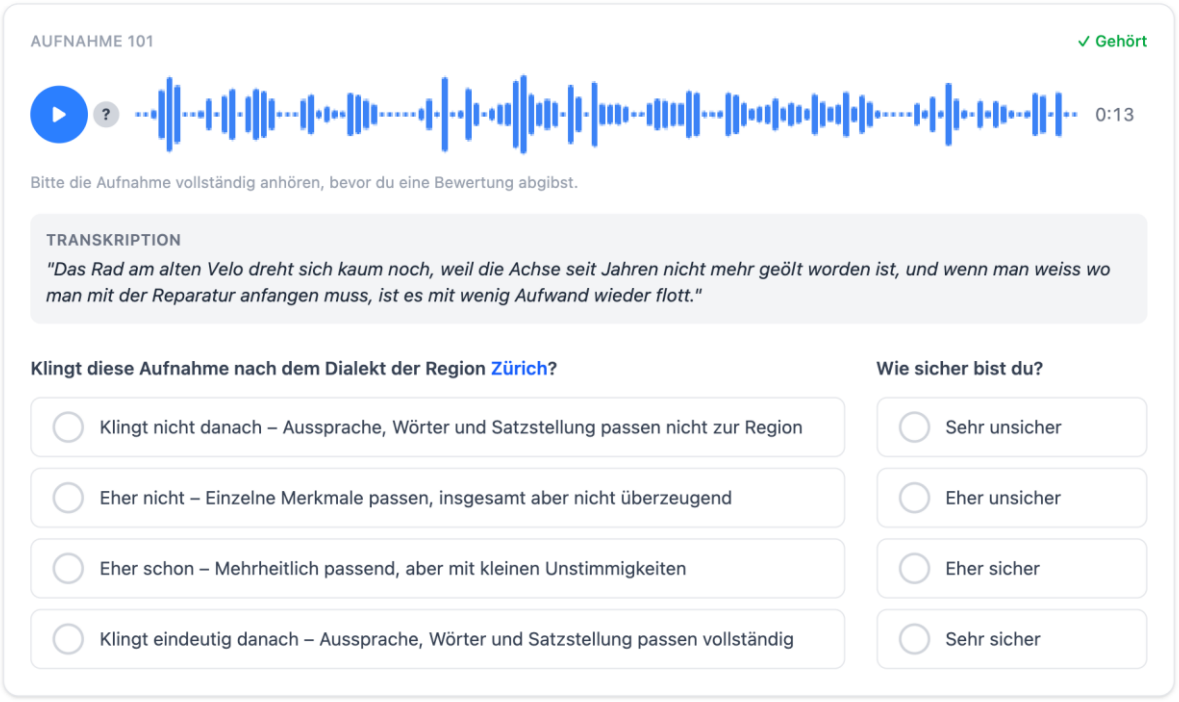

Abbildung 5: Annotationsmaske mit Audioplayer, Transkription sowie den Bewertungsskalen für Authentizität und Sicherheit. Oben befindet sich der Audioplayer zur Wiedergabe des Stimulus, gefolgt von der Texttranskription. Unten erfolgen zwei Eingaben: Links wird die Dialektauthentizität eingestuft, rechts die Annotationssicherheit auf einer vierstufigen Skala erfasst.

3.2 Auswahl der Referenzstimmen

Für das verwendete XTTS-Modell werden Referenzaufnahmen benötigt (*im Folgenden als Referenzstimme bezeichnet*). Da diese die direkte Grundlage für die Synthese bilden, ist ihre Qualität massgeblich für das Endresultat. Eine Grundvoraussetzung ist die Verwendung von vorgelesener Sprache *(Scripted Speech)*. Im Gegensatz zur Spontansprache, die durch unregelmässige Pausen, Füllwörter oder artikulatorische Störgeräusche geprägt ist, bietet vorgelesene Sprache eine hohe phonetische Konsistenz und eine kontrollierte Intonation. Die Auswahl der Referenzstimmen erfolgt über einen systematischen Selektionsprozess, bestehend aus automatisierter Vorfilterung und manueller Validierung.

Es werden Aufnahmen aus zwei etablierten Korpora herangezogen: dem «Thorsten-Voice Dataset 2022.10» [22] und dem «Common Voice Scripted Speech 24.0 – German»[2] [23]. Beide enthalten ausschliesslich vorgelesene Sprache, erfordern jedoch korpusspezifische Filter- und Bearbeitungsschritte:

- **Thorsten-Voice-Datensatz:** Da dieser Datensatz eher kurze Aufnahmen eines einzelnen professionellen Sprechers enthält, wurden zunächst alle Audiodateien mit einer Mindestdauer von 8 Sekunden gefiltert. Dies ergab ein Pool von 810 Kandidaten. Aufgrund der durchgehend hohen Audioqualität und Sprechdichte konnte die finale Auswahl direkt und ohne nachträgliche Bearbeitung getroffen werden.

[2] Für die Studie wurde «Common Voice Scripted Speech 24.0 – German» verwendet. Da dieser Datensatz zum Zeitpunkt der Berichterstellung nicht mehr verfügbar ist, wird auf die nachfolgende Version 25.0 verwiesen.

- **Common-Voice-Datensatz:** Diese Aufnahmen wiesen hingegen häufig längere Stillepausen auf. Um ausreichend nutzbare Sprachanteile für das Speaker Embedding zu gewährleisten, wurde hier eine höhere Mindestdauer von 14 Sekunden angesetzt, was 249 Kandidaten ergab. Diese durchliefen eine manuelle Qualitätskontrolle, bei der Aufnahmen mit Hintergrundgeräuschen, Audio-Glitches oder einem unpassenden Verhältnis von effektiver Sprechzeit zu Gesamtdauer aussortiert wurden. Die finale Auswahl wurde mit der Audiosoftware Audacity nachbearbeitet, um stille Abschnitte an den Rändern zu kürzen und störende Klickgeräusche zu entfernen.

Das Resultat dieses Selektionsprozesses ist ein Datensatz von fünf technisch und artikulatorisch einwandfreien Referenzstimmen:

- **speaker_01_male_de**: 3baa02facfb81b730c0ec59298008340.wav (*Thorsten-Voice*)
- **speaker_02_male_de**: common_voice_de_19683767.mp3 (*Common-Voice*)
- **speaker_03_male_de**: common_voice_de_43339826.mp3 (*Common-Voice*)
- **speaker_04_male_de**: common_voice_de_43345193.mp3 (*Common-Voice*)
- **speaker_05_female_de**: common_voice_de_43596423.mp3 (*Common-Voice*)

3.3 Auswahl der Sätze

Für die Qualität der Syntheseergebnisse ist das Design der Eingabesätze entscheidend. Die Zielvorgabe für die generierten Audioausgaben liegt bei einer Dauer von etwa 15 Sekunden. Diese Zeitspanne bildet einen Kompromiss zwischen evaluativer Genauigkeit und zeitlicher Effizienz: 15 Sekunden bieten ausreichend phonetisches Material für eine zuverlässige Dialekterkennung und halten gleichzeitig den Aufwand pro Annotation gering. Da in der Studie mehrere Dutzend Annotationen pro Person anfallen, ist diese Begrenzung erforderlich, um die Abbruchquoten aufgrund einer zu langen Befragung zu minimieren. Empirisch führen Textsegmente mit einem Umfang von etwa 200 Zeichen zu Synthesen dieser Ziellänge.

Um eine solche Textmenge natürlich zu gliedern, erschien die Verwendung von zwei kürzeren Sätzen pro Audioausgabe zunächst naheliegend. Vortests mit dem TTS-Modell zeigten jedoch, dass aufeinanderfolgende Sätze häufig ohne natürliche Pause als kontinuierlicher Redefluss artikuliert werden. Dies beeinträchtigt die Natürlichkeit der Synthese massgeblich, da die Satzmelodie (*Prosodie*) am Satzende nicht sauber abgeschlossen wird, sondern direkt in den Folgesatz übergeht. Daher wurden für den Datensatz ausschliesslich längere Einzelsätze konzipiert, die bewusst auf komplexe syntaktische Verschachtelungen verzichten.

Neben der Satzlänge beeinflusst auch die gewählte Zeitform die Synthesequalität. Da das Präteritum im Schweizerdeutschen fast vollständig fehlt, wird im Alltag primär das Perfekt genutzt [24]. Die Syntax innerhalb von Perfektkonstruktionen variiert jedoch stark zwischen den Dialektregionen (*z. B. Bern: «wüu er isch gschprunge» vs. Zürich: «wil er gschprunge isch»*). Um zu verhindern, dass solche strukturellen Unterschiede beim TTS-Mapping zu prosodischen Artefakten oder unnatürlichen Betonungen führen, wurden alle Sätze im Präsens oder in einer zeitunabhängigen Form verfasst.

Damit die generierten Sätze messbare und eindeutige dialektale Merkmale aufweisen, wurden gezielt phonetische Ankerpunkte integriert. Die Auswahl dieser Schlüsselwörter, die zur empirischen Differenzierung der Dialektregionen dienen, basiert auf dem Dialäktatlas [25]. Dieser vergleicht die raumzeitliche Verteilung der Aussprache über drei Generationen. Um den aktuellen Sprachstand abzubilden und der Zielgruppe der Annotierenden zu entsprechen, wurden die Daten der dritten Generation (*geboren 1985–2002, befragt 2020–2023*) herangezogen. Als

phonetische Grundlage dienen somit Wörter, die in der gesamten Deutschschweiz semantisch bekannt sind, regional jedoch unterschiedlich ausgesprochen werden.

Der Fokus liegt auf der phonetischen Variation (*Lautung*) statt auf der lexikalischen Variation (*Vokabular*). Die Konstruktion mit stark voneinander abweichenden regionalen Begriffen (*z. B. «Zuun» vs. «Haag» oder «Ghüder» vs. «Chübel»*) wurde verworfen. Initiale Tests zeigten, dass die untersuchten TTS-Modelle regionales Vokabular häufig nicht kennen und stattdessen auf eine hochdeutsche Standardaussprache ausweichen. Um die Dialektadaption gezielt zu messen, dienen daher ausschliesslich Wörter mit schweizweit identischem Wortstamm, aber regional distinkter Aussprache als Grundlage (*z. B. die Lautung von «Brille»*). Dies stellt sicher, dass die Systeme nicht an fehlendem Vokabular scheitern, sondern anhand der korrekten phonetischen Realisierung bewertet werden.

Initiale Rohentwürfe der Sätze wurden mithilfe des Large Language Models Claude Sonnet 4.5 generiert. Diese Entwürfe dienten als Ausgangslage und wurden manuell überarbeitet und gezielt um die phonetischen Ankerpunkte herum konstruiert, um die methodischen Vorgaben (*Satzstruktur, Zeitform und Ziellänge*) präzise zu erfüllen. Das Resultat sind fünf finale Sätze, in denen die dialektrelevanten Schlüsselwörter aus dem Dialäktatlas (*hier fett gedruckt hervorgehoben*) eingebettet sind:

- **utt_001**: Beim Kochen sollte man das **Küchenfenster** immer einen Spalt offen lassen, sonst zieht der Geruch stundenlang durch die ganze Wohnung, besonders dann, wenn man Fisch oder stark gewürzte Speisen zubereitet.
- **utt_002**: Ein guter **Käse** braucht Zeit, und wer ihn zu früh anschneidet, bereut es spätestens beim ersten Biss, weshalb man ihn am besten zuerst bei Raumtemperatur stehen lässt, bevor man ihn serviert.
- **utt_003**: Mit einer neuen **Brille** sieht man die Welt wieder scharf und die **Augen** müssen sich zuerst daran gewöhnen, besonders wenn man **tief** in die Ferne schaut und Details wahrnimmt, die man vorher nicht gesehen hat.
- **utt_004**: Von der alten **Brücke** aus hat man einen schönen Blick auf die Landschaft in der **Tiefe**, und viele Menschen bleiben deshalb kurz stehen, um die Aussicht zu bewundern und Fotos zu machen.
- **utt_005**: Das **Rad** am alten Velo dreht sich kaum noch, weil die **Achse** seit Jahren nicht mehr geölt worden ist, und wenn man weiss wo man mit der Reparatur anfangen muss, ist es mit wenig Aufwand wieder flott.

3.4 TTS-Pipeline und Qualitätssicherung

Wie in Kapitel 2.2 dargelegt, verläuft der Syntheseprozess von Zero-Shot-TTS-Modelle nicht deterministisch, was zu starken Qualitätsschwankungen der generierten Audioausgaben führt. Typische Probleme sind unvollständige Wortrealisierungen *(Truncation)* sowie akustische Artefakte wie Knackgeräusche oder digitales Rauschen. Um zu verhindern, dass solche Mängel die Annotationsstudie verfälschen, ist ein standardisiertes Qualitätssicherungsverfahrens unerlässlich. Da zunächst ein Vielfaches der benötigten TTS-Ausgaben generiert wird (*Iterationen*), um genügend fehlerfreie Kandidaten zu erhalten, übersteigt das Datenvolumen die Kapazitäten einer rein manuellen Überprüfung. Das Auswahlverfahren muss daher automatisiert und skalierbar ablaufen. Die technische Umsetzung wurde als modulare Code-Pipeline innerhalb einer Python-Umgebung realisiert und gliedert sich in drei sequenzielle Teilschritte, die in den nachfolgenden Unterkapiteln detaillierter beschrieben werden.

3.4.1 Generierung der Stimuli

Der erste Schritt umfasst die automatisierte Batch-Inferenz zur Synthese der TTS-Ausgaben auf Basis des XTTS-Frameworks (*siehe Kapitel 2.2*). Dabei wird das kartesische Produkt aus allen definierten Modellarchitekturen, Referenzstimmen, Sätzen und Dialekten berechnet. Um die zuvor beschriebenen Qualitätsschwankungen der Modelle auszugleichen, wird jede dieser Kombinationen mehrfach iteriert, um genügend fehlerfreien Kandidaten zu erhalten. Mit steigender Iterationsanzahl skaliert jedoch der Rechen- und Zeitaufwand für die gesamte Pipeline linear. Zur performanten und ausfallsicheren Skalierung nutzt die Architektur drei essenzielle Mechanismen:

- **Embedding-Caching:** Die rechenintensive Extraktion der Stimmmerkmale erfolgt nicht pro generierter TTS-Ausgabe, sondern initial einmal pro Referenzstimme. Diese Zwischenspeicherung reduziert die Inferenzzeit beim anschliessenden Zero-Shot Voice Cloning erheblich.
- **Explizites Speichermanagement:** Um eine Überlastung des VRAM der GPU bei grossen Batch-Durchläufen zu verhindern, wird nach der Abarbeitung eines Modells eine proaktive Speicherbereinigung durchgeführt.
- **Checkpointing und Metadaten-Tracking:** Jede erfolgreiche Synthese wird sofort in einem Metadatensatz protokolliert. Dies ermöglicht eine nahtlose Wiederaufnahme nach Systemunterbrechungen und verknüpft jede Audioausgabe mit ihren experimentellen Parametern anhand einer eindeutigen Identifikationsnummer (*UUID*).

3.4.2 Filterung der Stimuli

Im zweiten Schritt der Pipeline werden recheneffiziente, heuristische Audiomerkmale extrahiert, um die strukturelle Integrität der TTS-Ausgaben zu bewerten. Dabei stehen zwei akustische Basisindikatoren im Fokus: der durchschnittliche Signalpegel und die maximale Pausendauer.

Die durchschnittliche Lautstärke wird global über den quadratischen Mittelwert (*Root Mean Square, RMS*) der Amplitude des gesamten Audiosignals ermittelt. Ein ungewöhnlich niedriger RMS-Wert fungiert dabei als verlässlicher Indikator für systematische Synthesefehler, wie beispielsweise komplett stumme Audioausgaben, abrupte Abbrüche *(Truncation)* oder Signale mit unverhältnismässig leisen Segmenten.

Zusätzlich wird die Dauer der längsten zusammenhängenden Stillephase quantifiziert. Die Pausenerkennung erfolgt dabei dynamisch relativ zum Spitzenpegel der jeweiligen TTS-Ausgabe (*40 dB unterhalb des Maximums*). Dieser relative Ansatz macht den Algorithmus robust gegenüber Lautstärkeschwankungen zwischen den verschiedenen TTS-Modellen und Referenzstimmen. Überdurchschnittlich lange Pausen weisen auf eine fehlerhafte Prosodie oder ein vorzeitiges Abbrechen der Generierung hin.

Die extrahierten Indikatoren (*RMS-Wert und maximale Pausendauer*) werden direkt im zentralen Metadatensatz gespeichert. In dieser Phase findet noch keine aktive Eliminierung der Daten statt, jedoch dienen die Merkmale im nachfolgenden Bewertungsschritt als fundamentale Filterkriterien, um offensichtlich fehlerhafte Synthesen frühzeitig von der rechenintensiven Qualitätsbewertung auszuschliessen.

3.4.3 Automatische Bewertung der Stimuli

Im letzten Schritt der Pipeline erfolgt eine tiefgehende Qualitätsbewertung der generierten Synthesen. Bevor die rechenintensiven Metriken kalkuliert werden, greift an dieser Stelle die im vorherigen Kapitel definierten Filterkriterien: Es werden ausschliesslich TTS-Ausgaben zur Bewertung zugelassen, die eine RMS-Wert von ≥ 0.02 aufweisen und deren längste ununterbrochene Pause den Schwellenwert von 1.5 Sekunden nicht überschreitet. Ausgaben, die diese Minimalanforderungen nicht erfüllen, werden direkt vom finalen Datensatz ausgeschlossen.

Für die verbleibenden, validen Synthesen werden anschliessend in zwei sequenziellen Phasen der UTMOS sowie die WER berechnet (*siehe Kapitel 2.3*).

- **UTMOS:** Dieser Wert quantifiziert die wahrgenommene Sprachqualität auf einer kontinuierlichen Skala von 1 bis 5 (*höhere Werte stehen für bessere Qualität*) und wird mithilfe des Modells «utmos22_strong» aus dem Repository «tarepan/SpeechMOS» [26] ermittelt. Da dieses neuronale Netz spezifische Eingabeformate erfordert, werden die Audioausgaben vor der Inferenz auf eine Abtastrate von 16 kHz resampelt und in ein monophones Signal umgewandelt.
- **WER:** Die automatische Transkription zur Ermittlung der WER erfolgt mit dem Modell «Whisper large-v3» [27], wobei die Transkriptionssprache explizit auf Standarddeutsch fixiert wird. Da empirisch nachgewiesen ist, dass Whisper schweizerdeutsche Lautäusserungen robust in standarddeutsche Schrift überführt [28], ist ein direkter Abgleich mit dem standarddeutschen Referenzsatz aussagekräftig. Um zu verhindern, dass die Fehlerberechnung durch rein formale Abweichungen verfälscht wird, durchlaufen Referenztext und generiertes Transkript vorab eine linguistische Normalisierung: Alle Grossbuchstaben werden in Kleinbuchstaben umgewandelt und jegliche Interpunktion wird entfernt.

3.5 Modellauswahl

Stehen für eine experimentelle Studie mehrere Architektur- oder Checkpoint-Varianten zur Verfügung, ist ein systematischer Selektionsprozess erforderlich. Ziel dieses Schritts ist keine generelle Qualitätsbewertung der Modelle, sondern die Identifikation des für den spezifischen Anwendungsfall am besten geeigneten Modells. Für die vorliegende Arbeit wurden vier Modelle evaluiert, die im Rahmen einer vorausgehenden Masterarbeit trainiert und vom betreuenden Institut (*Centre for Artificial Intelligence, ZHAW*) für diesen Einsatzzweck vorselektiert wurden [3].

Um eine datengetriebene Entscheidung zu ermöglichen, wurden für jedes der vier Modelle 1600 TTS-Ausgaben generiert (*insgesamt 6400 Synthesen*). Die spezifische Parameterwahl resultiert aus einer methodischen Abwägung zwischen statistischer Signifikanz und Rechenaufwand:

- **8 Dialekte:** Diese setzen sich aus den sieben untersuchten Schweizer Dialektregionen sowie Standarddeutsch als linguistischer Kontrollgruppe (*Baseline*) zusammen.
- **10 Sätze:** Um die Stabilität über eine möglichst breite phonetische und syntaktische Varianz hinweg zu prüfen, wurde hier ein grösserer Satz-Pool genutzt. Für die spätere Annotationsstudie wurde diese Menge aus Gründen der zeitlichen Effizienz auf fünf Sätze reduziert.

- **2 Referenzstimmen:** Um die grundlegende Fähigkeit der Modelle zum Zero-Shot Voice Cloning zu verifizieren, genügt methodisch ein repräsentatives Subset von zwei Stimmen. Da das primäre Ziel der relative Leistungsvergleich ist, reicht diese Validierungsmenge, um qualitative Differenzen zwischen den Architekturen zuverlässig zu messen.
- **10 Iterationen:** Mehrfache Durchläufe kompensieren das nicht-deterministische Verhalten der End-to-End-Modelle. Sie stellen sicher, dass ausreichende statistische Datenpunkte vorliegen, um qualitätsschwankende Ausreisser über die spätere Median-Berechnung zuverlässig auszugleichen.

Nach der heuristischen Vorfilterung (*siehe Kapitel 3.4.2*) verblieben je nach Modell zwischen 1496 und 1589 verwertbare Audioausgaben (*N*), welche anschliessend die automatische Bewertungs-Pipeline (*UTMOS und WER*) durchliefen. Das automatisierte Selektionsverfahren operiert auf den aggregierten Scores in zwei Stufen:

1. **Primäres Kriterium (*UTMOS-Median*)**: In einem ersten Schritt werden die Modelle nach ihrem UTMOS-Median absteigend sortiert. Der Median dient als robustes Lagemass, um das Ranking nicht durch systembedingte, extreme Syntheseausreisser zu verzerren. Da der übergeordnete Verwendungszweck die Dialektbewertung durch menschliche Annotierenden ist, stehen eine natürliche Aussprache und Prosodie im Vordergrund. Eine Synthese, die roboterhaft klingt, ist für diesen Zweck unbrauchbar, selbst wenn der Text inhaltlich fehlerfrei realisiert wurde. Der UTMOS fungiert somit als relevanteste Qualitätsmetrik.
2. **Sekundäres Kriterium (*WER-Median*):** Bei einem allfälligen Gleichstand oder sehr ähnlichen UTMOS-Werten dient der WER-Median in aufsteigender Reihenfolge als sekundäres Sortierkriterium *(Tie-Breaker)*. Eine direkte mathematische Aggregation beider Metriken zu einem einzelnen Score ist methodisch nicht zielführend, da sie auf unterschiedlichen Skalen operieren und gegenläufige Optima aufweisen.

Gemäss diesem Verfahren lieferte das Modell mit der ursprünglichen Checkpoint-Bezeichnung *GPT_XTTS_v2.0_SwissGPC_epoch_6_subset_4* die besten Resultate. Wie in Tabelle 1 hervorgeht, erzielte es den höchsten UTMOS-Median bei gleichzeitig niedrigstem WER-Median. Der Übersichtlichkeit halber wird dieses Modell im weiteren Verlauf der Arbeit als «SwissGPC» bezeichnet.

Tabelle 1: Modell-Ranking basierend auf UTMOS- und WER-Medianwerten. Die Modelle sind absteigend nach Rangplatz sortiert und weisen die Anzahl der evaluierten Stimuli (N) aus. Ein höherer UTMOS steht für eine höhere wahrgenommene Sprachqualität; eine tiefere WER zeigt eine präzisere Textwiedergabe an.

Rang	Modell	N	UTMOS (Median)	WER (Median)
1	SwissGPC	1589	2.856	0.100
2	SRG+STT4SG	1586	2.749	0.103
3	SRG+STT4SG++	1533	2.508	0.103
4	Baseline (STT4SG only)	1496	2.262	0.139

Zur detaillierten Veranschaulichung der statistischen Verteilung sowie der jeweiligen Mediane pro Modell sind die UTMOS- und WER-Werte in Abbildung 6 und Abbildung 7 als Boxplots visualisiert. Auf Basis dieser eindeutigen, datengetriebenen Erkenntnisse wurde das Modell *SwissGPC* für die Generierung des finalen Datensatzes verwendet.

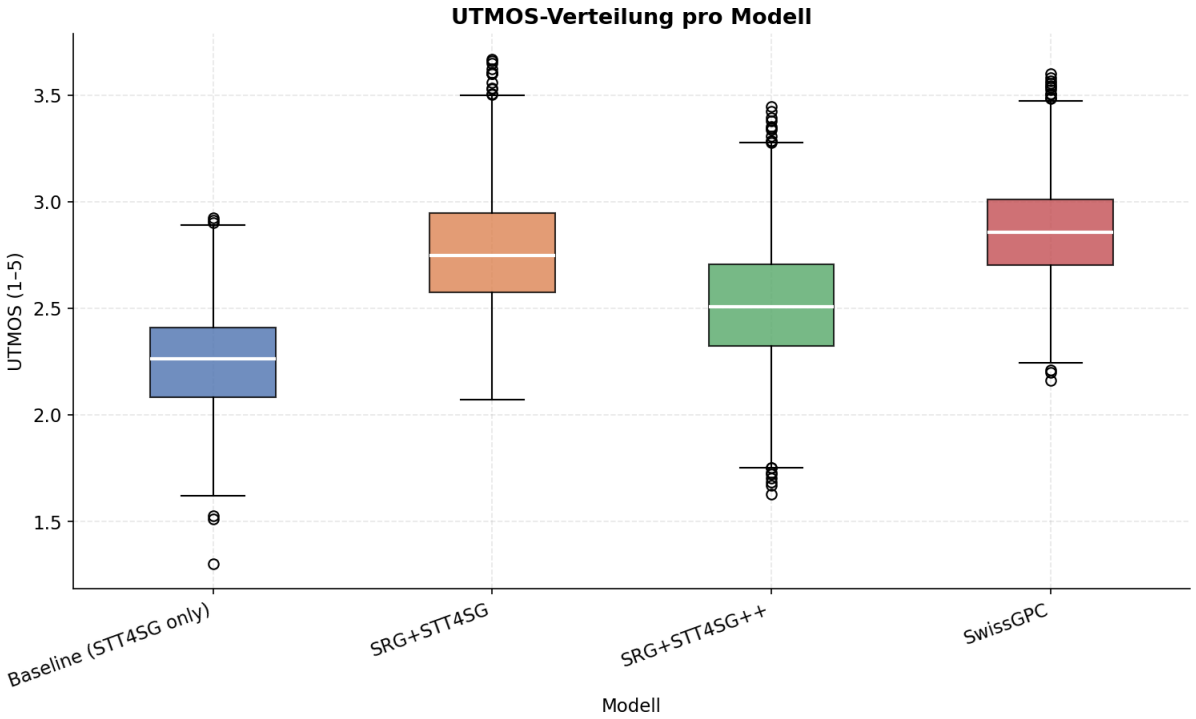

Abbildung 6: UTMOS-Verteilung pro Modell. Die Boxplots zeigen die Verteilung (Skala 1 bis 5) für die vier evaluierten Modellkonfigurationen. Dargestellt sind Median (horizontale Linie), Interquartilsabstand (Box) sowie Streuung und Ausreisser (Antennen).

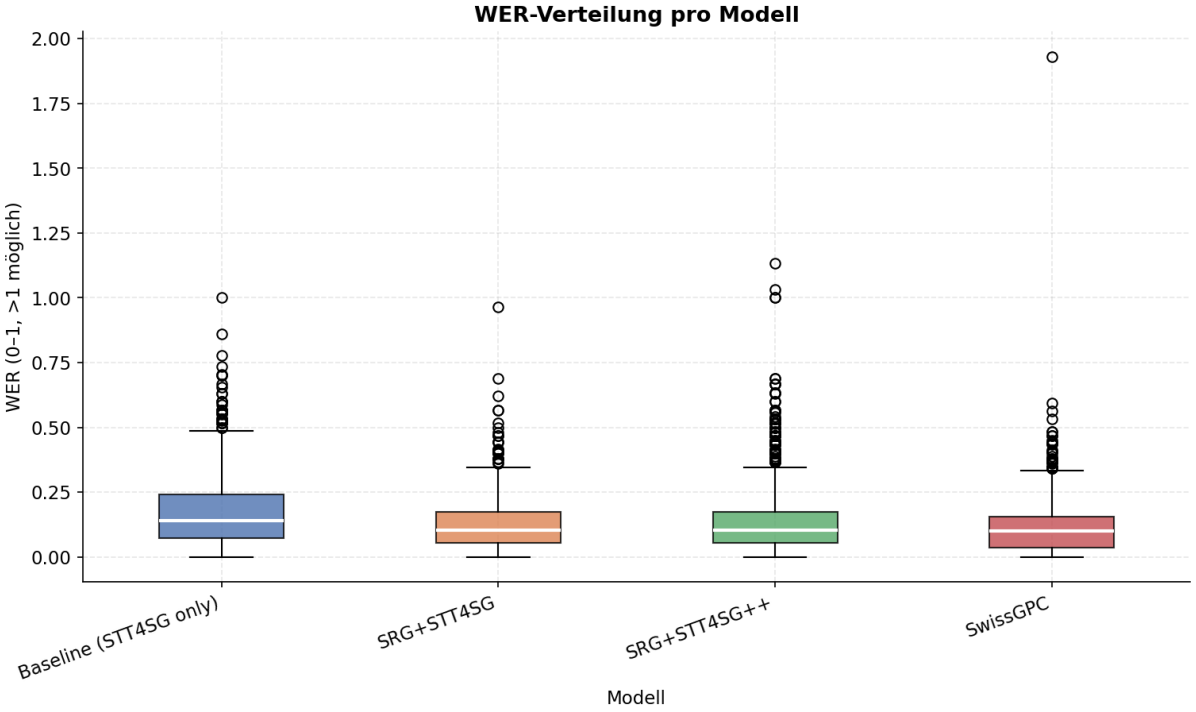

Abbildung 7: WER-Verteilung pro Modell. Die Boxplots zeigen die Verteilung der Word Error Rate (WER) für die vier evaluierten Modellkonfigurationen. Dargestellt sind Median (horizontale Linie), Interquartilsabstand (Box) sowie Streuung und Ausreisser (Antennen).

3.6 Generierung und Analyse des finalen Datensatzes

Die finalen Stimuli für die Annotationsstudie wurden mit dem ausgewählten Modell SwissGPC und der in Kapitel 3.4 etablierten Pipeline generiert. Um einen aussagekräftigen, aber für die menschlichen Annotierenden handhabbaren Datensatz zu erzeugen, wurde eine spezifische Parameterkombination definiert. Diese Wahl resultiert aus einer methodischen Abwägung zwischen linguistischer Repräsentativität und kognitiver Belastung während der Studie:

1. **7 Schweizerdeutsche Dialektregionen:** Diese Dimension ist durch die Trainingsdaten der evaluierten Modelle (*inklusive SwissGPC*) fest vorgegeben.
2. **5 Referenzstimmen:** Diese Anzahl bietet eine breite stimmliche Varianz. Damit wird sichergestellt, dass die Dialektqualität des Modells nicht nur bei einer spezifischen Referenzstimme funktioniert, sondern sich das Zero-Shot Voice Cloning robust verallgemeinern lässt.
3. **5 Sätze:** Die Satzanzahl wurde gegenüber der initialen Modellevaluation bewusst halbiert, um den zeitlichen Aufwand für die Teilnehmenden in einem realistischen Rahmen zu halten und die Gesamtmenge der zu bewertenden TTS-Ausgaben zu limitieren.
4. **10 Iterationen pro Kombination:** Wie bereits bei der Modellevaluation dient dies dem Ausgleich von generativen Qualitätsschwankungen. Aus den zehn generierten Versionen wird im Anschluss deterministisch jene Ausgabe mit dem höchsten UTMOS-Wert in den finalen Datensatz übernommen.

Dieses Setup ergibt zunächst 1750 generierte Rohsynthesen. Nach der Selektion der qualitativ besten Iteration ergibt sich ein finaler Datensatz von 175 Stimuli für die Annotationsstudie.

Für die praktische Durchführung der Datenerhebung wurde ein Zielwert von mindestens 50 Annotationen pro teilnehmende Person festgelegt (*geschätzter Zeitaufwand: ca. 30 Minuten*). Um die Teilnehmenden zum Erreichen dieses Ziels zu motivieren, wurde daran die Teilnahme an einer Verlosung von Gutscheinen geknüpft. Dieser Rahmen wurde pragmatisch gewählt, um einerseits eine ausreichende Menge an Datenpunkten pro Person zu generieren, andererseits aber die Durchführungsdauer niedrig zu halten. Den Teilnehmenden stand es frei, weniger als 50 Annotationen abzugeben oder auch den ganzen Datensatz zu annotieren.

Statistische Qualitätsverteilung des Datensatzes
Um eventuelle systematische Einflüsse auf die spätere menschliche Bewertung bereits im Vorfeld identifizieren zu können, wurde die Qualitätsverteilung (*WER und UTMOS*) der 175 finalen Stimuli in Bezug auf Dialekt, Referenzstimme und Satz statistisch analysiert.

Abbildung 8 visualisiert die Verteilung nach Dialekten. Die Qualitätsmetriken verhalten sich über die meisten Regionen hinweg konsistent. Eine signifikante Ausnahme bildet der Walliserdialekt (*VS*), welcher eine merklich höhere WER-Streuung aufweist. Dies deutet darauf hin, dass dieser phonetisch und lexikalisch stark abweichende Dialekt das Modell bei der fehlerfreien Wortsynthese vor überdurchschnittliche Herausforderungen stellt.

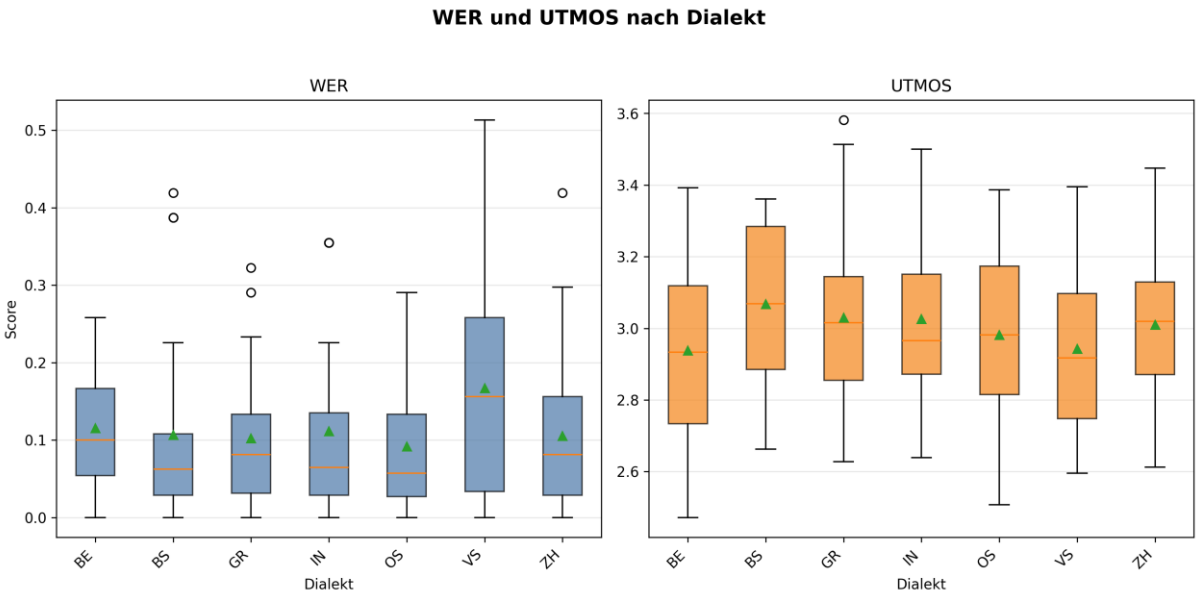

Abbildung 8: WER- und UTMOS-Verteilung nach Dialekt. Aufschlüsselung der Stimuli nach den sieben untersuchten Dialekten (BE, BS, GR, IN, OS, VS, ZH). Die Boxplots zeigen Median (orange Linie), Mittelwert (grüner Pfeil), Interquartilsabstand (Box) sowie Streuung und Ausreisser (Antennen).

Bei der Analyse nach der Referenzstimme (*siehe Abbildung 9*) zeigt sich, dass die WER zwischen den Profilen nur marginal variiert. Im Gegensatz dazu treten beim UTMOS deutliche Diskrepanzen auf: Während das Stimmprofil *speaker_03_male_de* überdurchschnittlich hohe Scores erzielt, fallen die Synthesen der Profile *speaker_04_male_de* und *speaker_05_female_de* in ihrer akustischen Qualität ab.

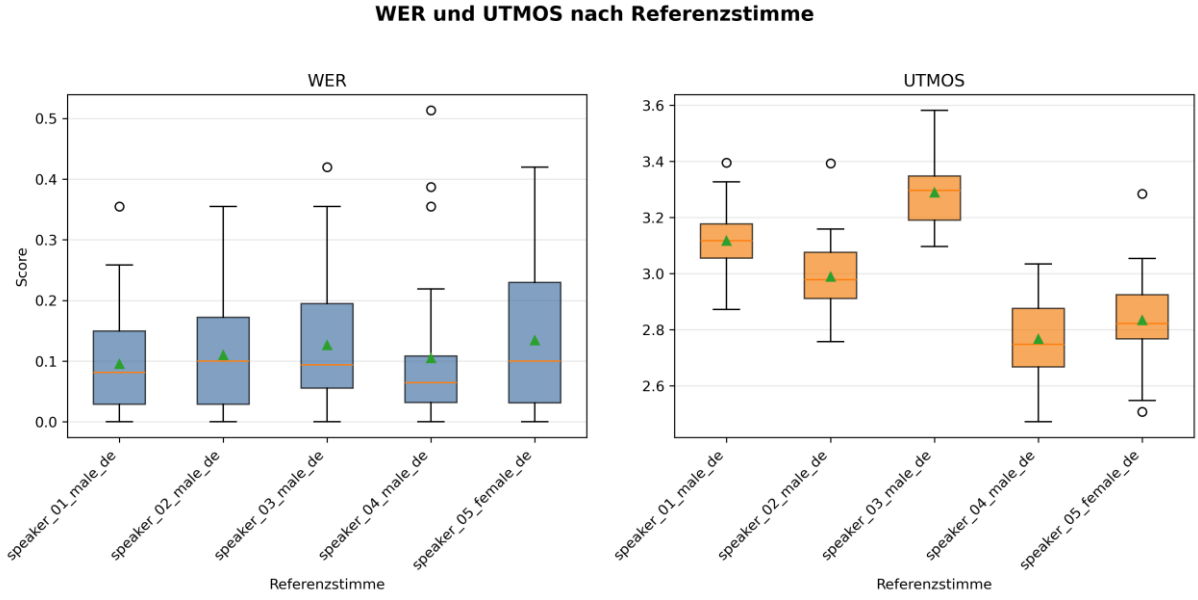

Abbildung 9: WER- und UTMOS-Verteilung nach Referenzstimme. Aufschlüsselung der Stimuli nach den fünf Referenzstimmen (speaker_01 bis speaker_05). Die Boxplots zeigen Median (orange Linie), Mittelwert (grüner Pfeil), Interquartilsabstand (Box) sowie Streuung und Ausreisser (Antennen).

Die Betrachtung der einzelnen Sätze (*siehe Abbildung 10*) zeigt ebenfalls ein weitgehend homogenes Bild, mit einer prägnanten Ausnahme: Satz 2 (*utt_002*) weist im Vergleich zu den restlichen Sätzen eine drastisch verschlechterte WER-Verteilung auf. Es ist davon auszugehen,

dass dieser spezifische Satz linguistische, syntaktische oder phonetische Strukturen aufweist, welche die Modellarchitektur systematisch zu fehlerhaften Synthesen zwingen.

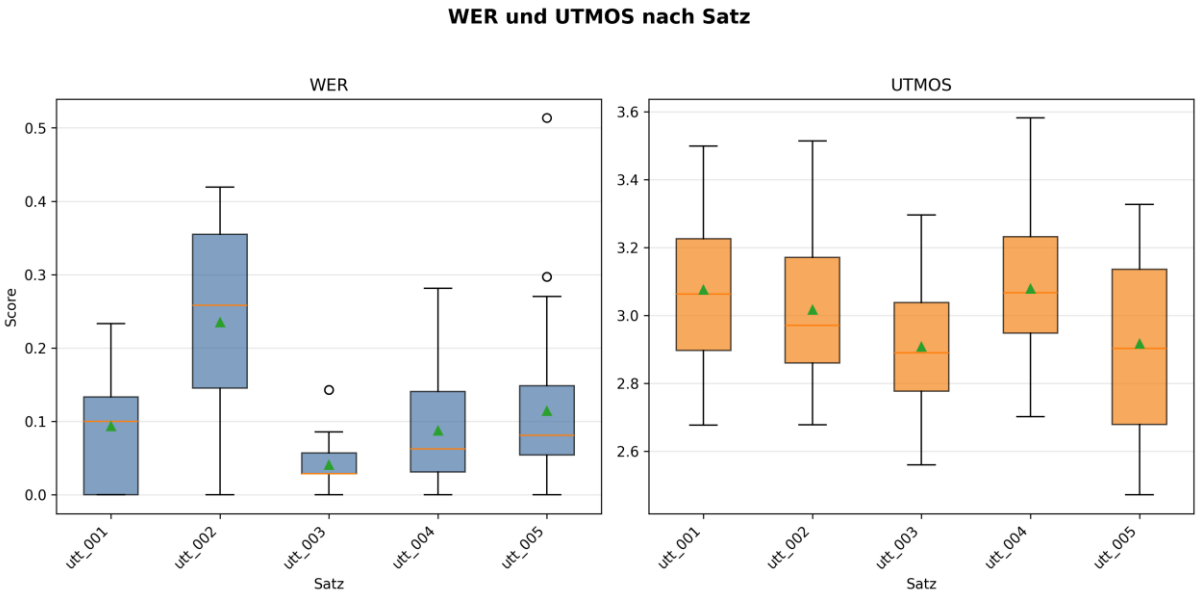

Abbildung 10: WER- und UTMOS-Verteilung nach Satz. Aufschlüsselung der Stimuli nach den fünf Sätzen (utt_001 bis utt_005). Die Boxplots zeigen Median (orange Linie), Mittelwert (grüner Pfeil), Interquartilsabstand (Box) sowie Streuung und Ausreisser (Antennen).

Abschliessend ist festzuhalten, dass die in diesem Kapitel präsentierten objektiven Metriken (*WER und UTMOS*) ausschliesslich deskriptiven Zwecken dienen. Sie belegen die strukturelle Qualität und die Verteilungseigenschaften des finalen Datensatzes vor Beginn der eigentlichen Datenerhebung. Da der methodische Fokus der nachfolgenden Untersuchung auf der statistischen Evaluation der subjektiven menschlichen Annotationen liegt, werden diese maschinellen Vorab-Scores im weiteren Verlauf der Arbeit nicht weiterführend analysiert oder mit den erhobenen Annotationsdaten korreliert.

4 Annotationsstudie

In diesem Kapitel werden die im Rahmen der Annotationsstudie erhobenen Daten deskriptiv aufbereitet. Zunächst erfolgt eine Beschreibung der Teilnehmenden hinsichtlich ihrer demografischen Merkmale und ihrer Selbsteinschätzung zur Dialekterkennung. Darauf aufbauend werden die Resultate der Kalibrierungsphase analysiert sowie der resultierende Annotationsdatensatz bezüglich Umfang und Verteilung vorgestellt. Die tiefergehende inhaltliche Auswertung und Interpretation der eigentlichen Annotationen erfolgen separat in Kapitel 6.

4.1 Teilnehmende

Die Analyse der demografischen Zusammensetzung sowie der Selbsteinschätzung der Teilnehmenden bildet eine wesentliche Grundlage für die spätere Interpretation der Studienergebnisse.

An der Annotationsstudie (*23. April bis 21. Mai 2026*) nahmen 195 Personen teil (*siehe Abbildung 11*). Die Rekrutierung erfolgte über eine ZHAW-interne Mailingliste, Adresslisten der Dialektsammlung sowie private Kontakte der Autoren. Im Verlauf des Versuchsprozesses reduzierte sich diese Stichprobe wie folgt:

- **175 Personen** schlossen die die Kalibrierungsphase vollständig ab.
- **168 Personen** erfassten im Anschluss mindestens eine verwertbare Annotation.
- **98 Personen** erreichten das in der Methodik definierte Basisziel von 50 oder mehr Annotationen.

Damit hat rund die Hälfte (*50 %*) der ursprünglichen Gesamtstichprobe die angestrebte Zielgrösse erfüllt oder überschritten.

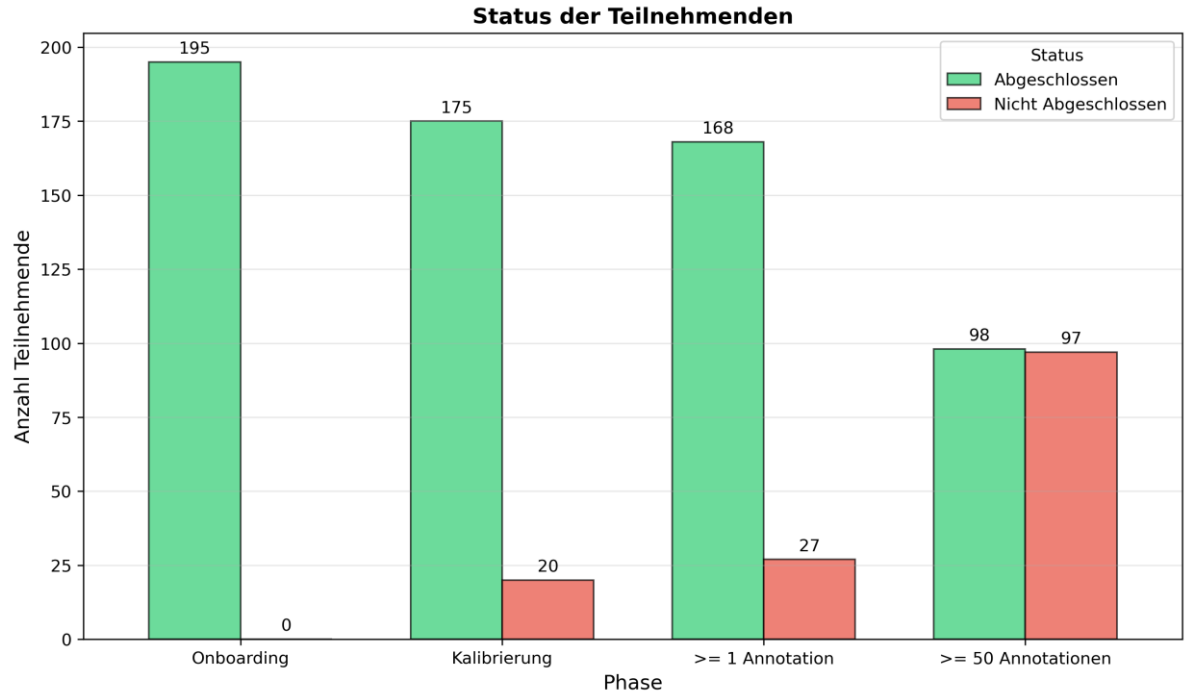

Abbildung 11: Status der Teilnehmenden nach Studienphase. Darstellung der Teilnehmenden (Y-Achse) über die vier Studienphasen (X-Achse). Grün markiert den erfolgreichen Abschluss einer Phase, Rot kennzeichnet Abbrüche. Die exakten Werte stehen über den Balken.

Die Verteilung der geleisteten Annotationen pro Kopf zeigt einen Median von exakt 51 Annotationen und entspricht damit dem angestrebten Zielkorridor (*siehe Abbildung 12*). Besonders hervorzuheben ist dabei, dass eine Kerngruppe von 27 Personen den vollständigen Datensatz lückenlos annotiert hat.

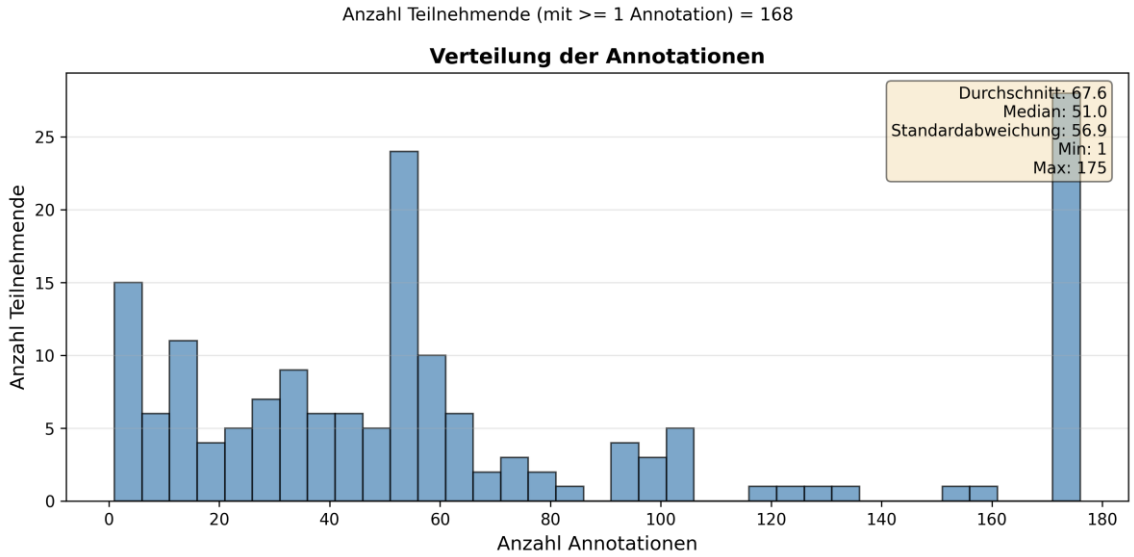

Abbildung 12: Verteilung der Annotationen pro teilnehmende Person. Das Histogramm zeigt die abgegebenen Annotationen (X-Achse, Spanne 1 bis 175) bezogen auf die Anzahl der Teilnehmenden (Y-Achse). Die Stichprobe umfasst Personen mit mindestens einer Annotation.

Hinsichtlich der Altersstruktur liegt der Schwerpunkt in der Altersgruppe zwischen 18 und 35 Jahren (*siehe Abbildung 13*). Bei der Selbsteinschätzung der eigenen Dialektregion zeigt sich zudem eine deutliche geografische Asymmetrie (*siehe Abbildung 14*): Die Dialektregion Zürich (*ZH*) ist im Vergleich zum restlichen Sprachraum stark überrepräsentiert, während die Regionen Basel, Graubünden und das Wallis unterrepräsentiert sind. Personen, die sich als nicht dialektsprechend einstufen, sind im Datensatz kaum vertreten. Diese starke Zentrierung auf den Raum Zürich spiegelt sich konsistent in den angegebenen Aufenthaltsorten der Teilnehmenden wider (*siehe Abbildung 15*).

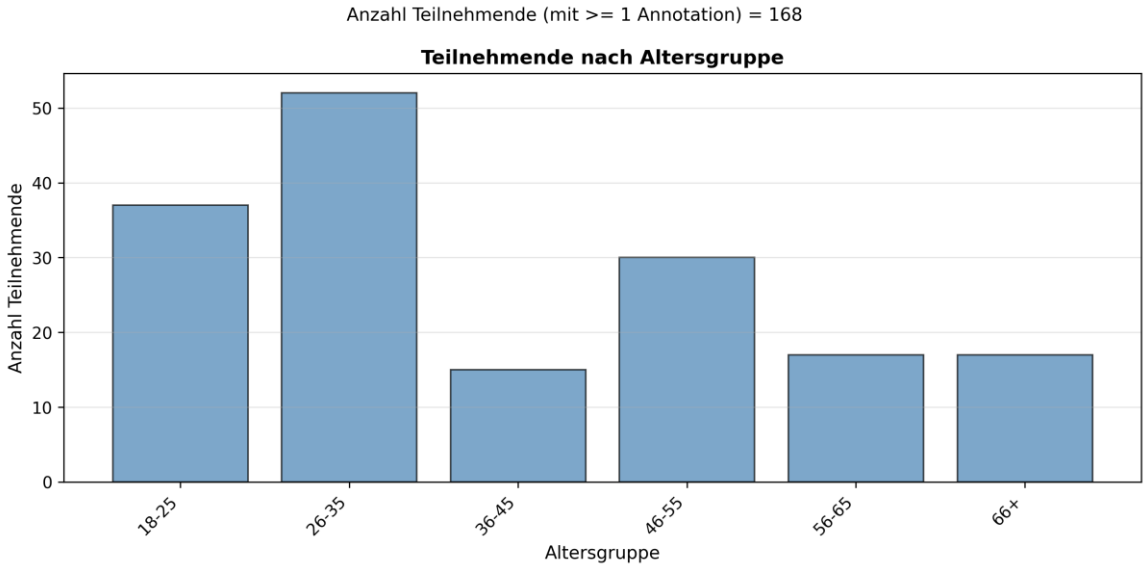

Abbildung 13: Verteilung der Teilnehmenden nach Altersgruppe. Anzahl der Teilnehmenden (Y-Achse) aufgeteilt nach den sechs Altersgruppen (X-Achse). Die Stichprobe umfasst Personen mit mindestens einer Annotation.

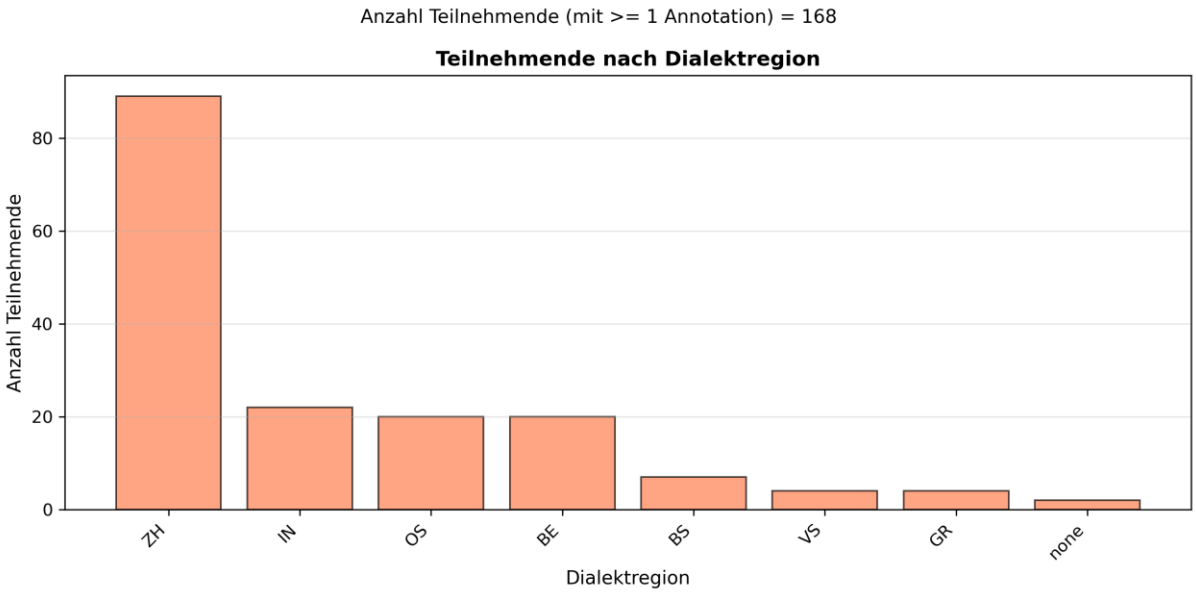

Abbildung 14: Verteilung der Teilnehmenden nach Dialektregion. Anzahl der Teilnehmenden (Y-Achse) nach selbst angegebener Dialektregion (X-Achse). Die Kategorie «none» erfasst Personen ohne Zuordnung zu einer spezifischen Schweizer Region. Die Stichprobe umfasst Personen mit mindestens einer Annotation.

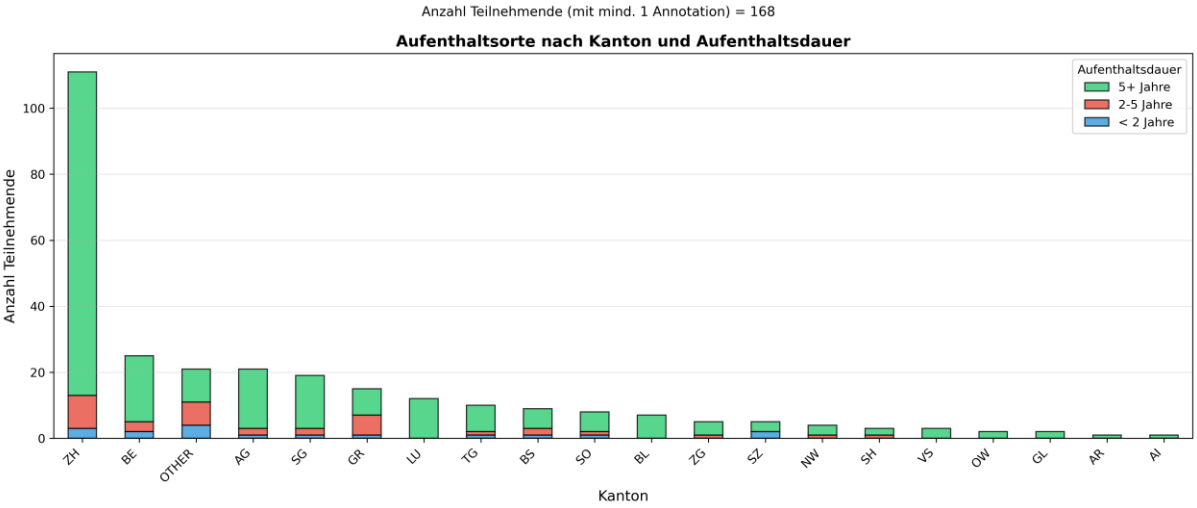

Abbildung 15: Aufenthaltsorte der Teilnehmenden nach Kanton und Aufenthaltsdauer. Anzahl der Teilnehmenden (Y-Achse) pro Kanton (X-Achse). Die gestapelten Säulen segmentieren nach Aufenthaltsdauer: über 5 Jahre (grün), 2 bis 5 Jahre (rot) und unter 2 Jahre (blau). «OTHER» fasst übrige Kantone und Auslandsaufenthalte zusammen.

Bezüglich der subjektiven Kompetenz bei der Dialekterkennung schätzen die Teilnehmenden ihre eigenen Fähigkeiten für die meisten Regionen insgesamt als sehr gut ein (*siehe Abbildung 16*). Eine markante Ausnahme bildet hierbei der Innerschweizer Dialekt (*IN*), dessen auditive Merkmale im Quervergleich systematisch als schwerer erkennbar eingestuft werden.

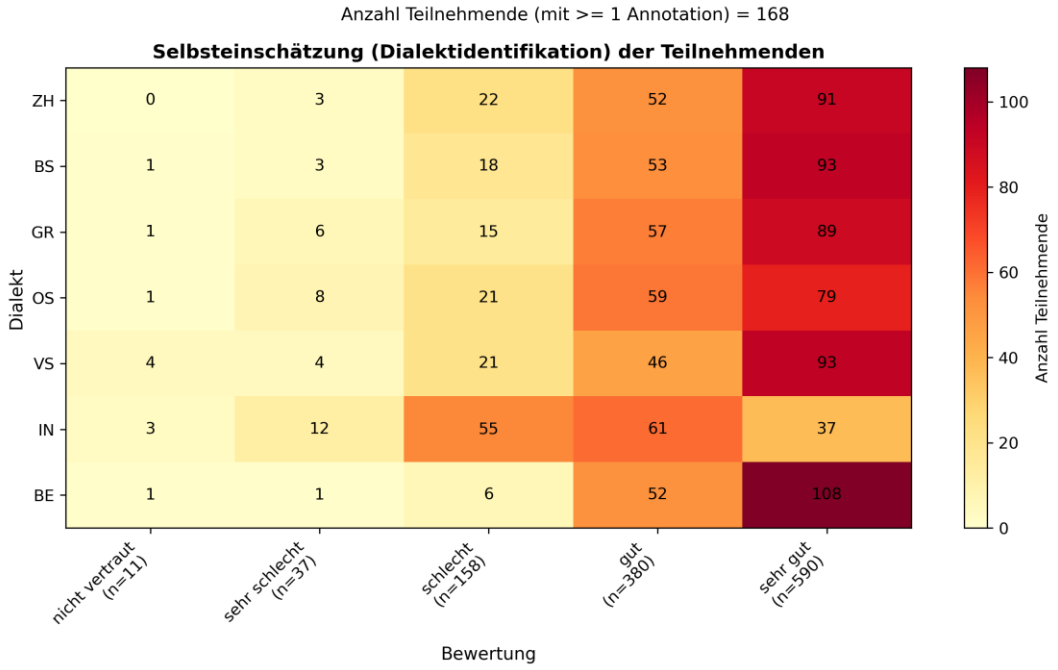

Abbildung 16: Selbsteinschätzung (Dialektidentifikation) der Teilnehmenden. Die Heatmap kreuzt die sieben Dialekte (Y-Achse) mit der fünfstufigen Likert-Skala («nicht vertraut» bis «sehr gut», X-Achse). Zellenwerte und Farbintensität (Gelb bis Dunkelrot) zeigen die Nennungshäufigkeit. Klammerwerte (X-Achse) summieren die Bewertungen pro Stufe.

Zusammenfassend ergibt sich für die demografische Struktur der Studie ein klares Bild: Teilnehmende mit einer Zürcher Dialektaffinität sind deutlich überrepräsentiert. Diese strukturelle Verzerrung muss bei den nachfolgenden Auswertungen und der Interpretation der Annotationsqualität zwingend als Störvariable berücksichtigt werden.

4.2 Kalibrierung

Dieser Abschnitt dokumentiert die Ergebnisse der Kalibrierungsphase der Annotationsstudie. Neben den erzielten Kalibrierungswerten und deren Zusammenhang mit der Selbsteinschätzung der Teilnehmenden wird die Eignung der einzelnen Kalibrierungsaufnahmen zur Differenzierung der Dialekterkennungskompetenz evaluiert. Darüber hinaus erfolgt eine Korrelationsanalyse zwischen Sicherheit und Korrektheit sowie eine Untersuchung hinsichtlich eines potenziellen Eingewöhnungseffekts im Verlauf der Sequenz.

4.2.1 Kalibrierungsresultate

Die Verteilung der Kalibrierungsresultate zeigt deutliche Unterschiede zwischen den einzelnen Regionen (*siehe Abbildung 17*). Die Dialekte aus Basel, Bern, dem Wallis und der Ostschweiz erzielten im Median höhere Scores und wurden somit von den Teilnehmenden präziser identifiziert. Im Gegensatz dazu weist die Region Graubünden einen negativen Median auf. Diese systematische Fehlklassifikation lässt vermuten, dass die Region Graubünden in den verwendeten Beispielen entweder untypische Merkmale aufwies oder häufig mit sprachlich nahestehenden Nachbarregionen verwechselt wurde.

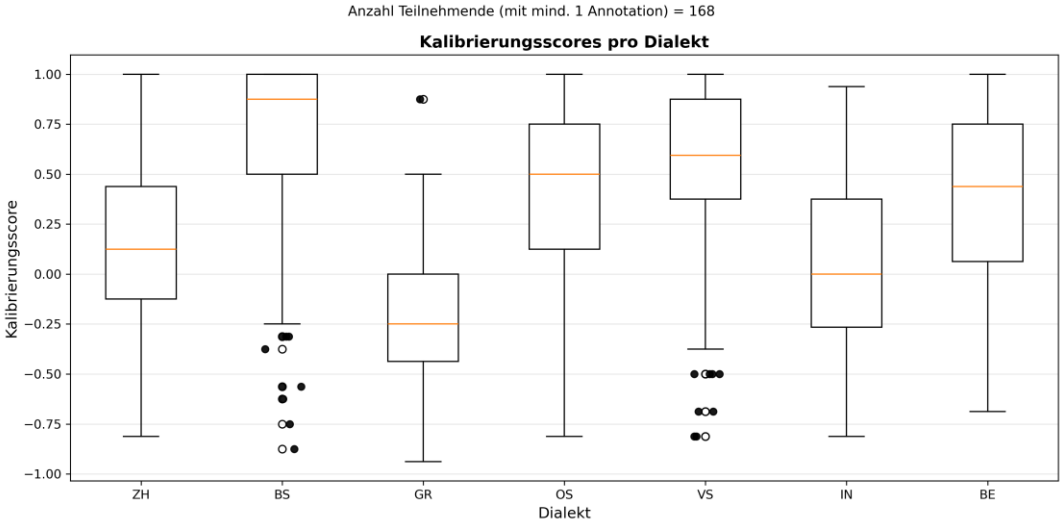

Abbildung 17: Kalibrierungsscore pro Dialekt. Verteilung der erreichten Kalibrierungsscores (Y-Achse, -1.00 bis 1.00) für die sieben evaluierten Schweizer Dialekte (X-Achse). Die Boxplots zeigen Median (orange Linie), Interquartilsabstand (Box) sowie Streuung und Ausreisser (Punkte ausserhalb der Antennen).

Der Zusammenhang zwischen der subjektiven Selbsteinschätzung der Teilnehmenden und ihren tatsächlichen Kalibrierungsscores verdeutlicht diesen Trend (*siehe Abbildung 18*). Aufgrund einer zu geringen Datenbasis wurde die Kategorie „nicht vertraut“ in dieser Auswertung nicht berücksichtigt. Generell zeigt sich für fast alle Dialektregionen der erwartete positive Zusammenhang: Je sicherer sich die Teilnehmenden im Vorfeld einschätzten, desto höher fällt ihr tatsächlicher Score aus.

Die starken Schwankungen und untypischen Mediane in der Gruppe «sehr schlecht» sind rein methodisch bedingt, da die geringe Datendichte in diesem Segment keine verlässliche Interpretation erlaubt. Als auffällige Anomalie sticht einzig die Dialektregion Graubünden hervor, bei der diese positive Korrelation komplett ausbleibt. Diese spezifische Problematik wird im folgenden Abschnitt qualitativ genauer untersucht.

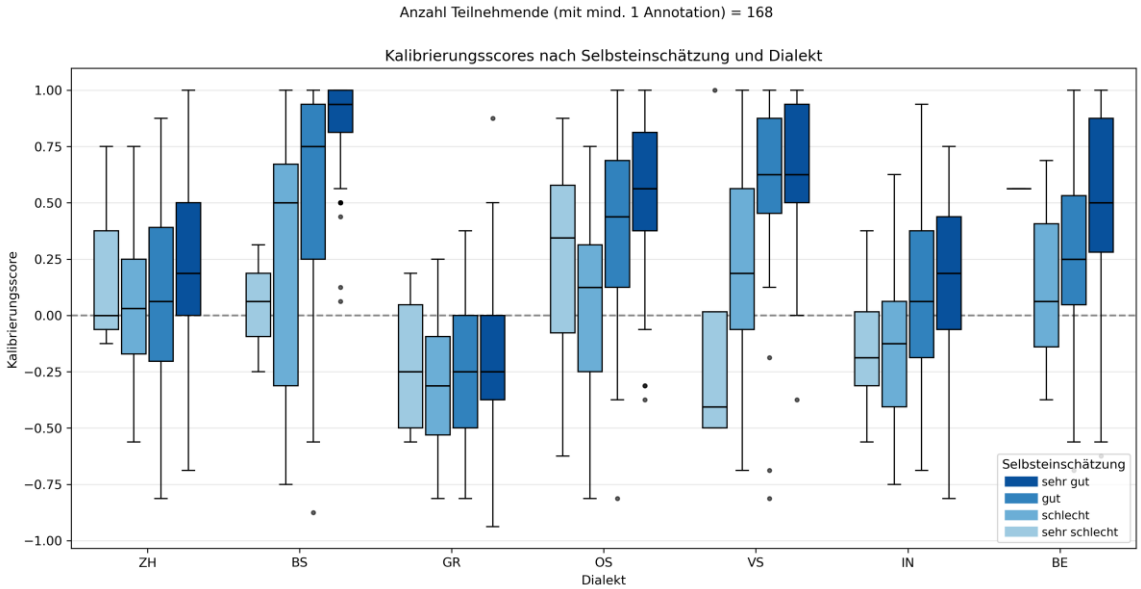

Abbildung 18: Zusammenhang Kalibrierungsscores und Selbsteinschätzung. Verteilung der Kalibrierungsscores (Y-Achse) für die sieben Dialektregionen (X-Achse). Die Farben differenzieren die Teilnehmenden nach ihrer subjektiven Selbsteinschätzung («sehr schlecht» bis «sehr gut»). Die Boxplots zeigen Median (schwarze Linie), Interquartilsabstand (Box) sowie Streuung und Ausreisser (Punkte ausserhalb der Antennen).

4.2.2 Eignung der Kalibrierungsaufnahmen

Um die Eignung der 28 Kalibrierungsaufnahmen zur Einschätzung der Dialekterkennungskompetenz zu evaluieren, wurden diese anhand ihrer durchschnittlichen Sicherheit und Korrektheit in vier Quadranten eingeteilt (*siehe Abbildung 19*):

- **Oben rechts (*Hohe Sicherheit / hohe Korrektheit*):** Diese Aufnahmen wurden von der Mehrheit der Teilnehmenden sicher und korrekt erkannt.
- **Oben links (*Tiefe Sicherheit / hohe Korrektheit*):** Die korrekte Zuordnung geht mit grosser Unsicherheit einher, was methodisch auf zufällige Treffer hindeuten kann.
- **Unten rechts (*Hohe Sicherheit / tiefe Korrektheit*):** Diese Aufnahmen bergen das Risiko einer systematischen Verzerrung, da die Teilnehmenden hier mit hoher Überzeugung einen falschen Dialekt gewählt haben (systematischer Fehler).
- **Unten links (*Tiefe Sicherheit / tiefe Korrektheit*):** In diesem Quadranten lässt sich die Fehlklassifikation durch die bewusste Unsicherheit der Teilnehmenden erklären.

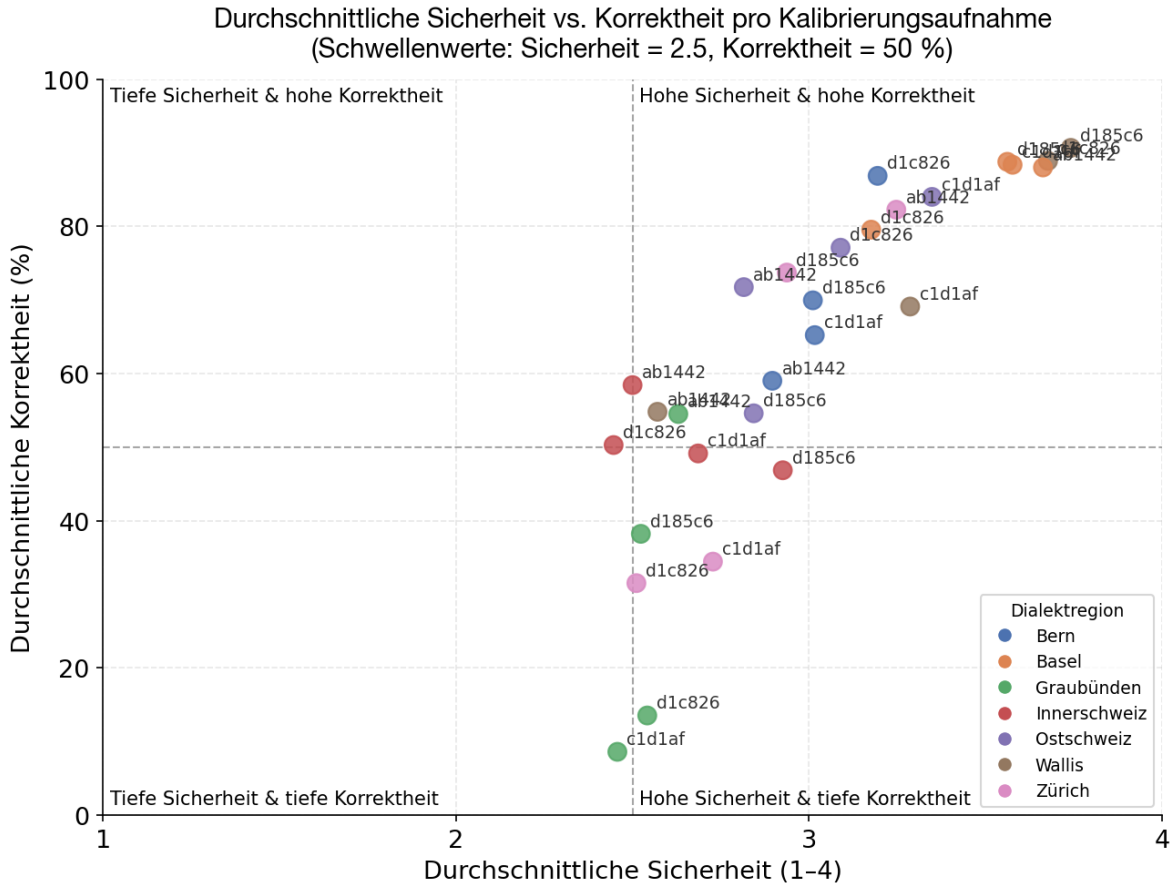

Abbildung 19: Durchschnittliche Sicherheit vs. Korrektheit pro Kalibrierungsaufnahme. Verteilung der Kalibrierungsstimuli (beschriftet nach IDs) anhand der mittleren Annotationssicherheit (X-Achse) und Korrektheit (Y-Achse). Gestrichelte Linien markieren Schwellenwerte (Sicherheit = 2.5; Korrektheit = 50 %) und bilden vier Kompetenzquadranten. Die Farben repräsentieren die Dialektregionen.

Der Grossteil der 28 Kalibrierungsaufnahmen befindet sich im Quadranten für hohe Sicherheit und hohe Korrektheit, was die grundsätzliche Validität des Kalibrierungsdatensatzes stützt. Ausserdem ist eine generelle Tendenz zur Sicherheit bei den Teilnehmenden festzustellen: Nur bei zwei der 28 Aufnahmen liegt die durchschnittliche Sicherheit unter dem Schwellenwert von 2.5.

Unabhängig von dieser allgemeinen Sicherheitstendenz erfordern zwei spezifische Aufnahmen eine qualitative Detailbetrachtung, da sie markante Abweichungen in der Erkennungsrate aufweisen. Beide betreffen den Dialekt Graubünden und verzeichnen die tiefsten Korrektheitsraten des gesamten Datensatzes.

Die Aufnahme *c1d1af* erreicht bei einer durchschnittlichen Sicherheit von 2.45 eine durchschnittlich Korrektheit von lediglich 8.6 %. Eine qualitative Prüfung zeigt, dass die Audioqualität zwar einwandfrei ist, der Dialekt jedoch akustisch nicht eindeutig identifizierbar ist. Als untypische Merkmale fallen ein überdurchschnittlich hohes Sprechtempo sowie ein leichtes Stocken im Redefluss auf.

Ein ähnliches Bild zeigt sich bei der Aufnahme *d1c826*, die mit 13.5 % (*bei einer durchschnittlichen Sicherheit von 2.54*) die zweittiefste durchschnittliche Korrektheit verzeichnet. Die Tonqualität weist hier ein leichtes Rauschen auf. Wie beim ersten Beispiel fehlen klare phonetische Merkmale für eine eindeutige Zuordnung, und es ist eine vergleichbare Verzögerung durch einen Neuansatz beim Lesen hörbar.

Zusammenfassend lässt die qualitative Detailbetrachtung darauf schliessen, dass weder die Audioqualität noch die festgestellten Unregelmässigkeiten im Redefluss die primäre Ursache für die tiefe Korrektheitsrate darstellen. Vielmehr scheinen die charakteristischen Merkmale des Graubündner Dialekts in diesen spezifischen Aufnahmen nicht prägnant genug ausgeprägt zu sein, was die korrekte Zuordnung durch die Teilnehmenden massgeblich erschwerte.

Ergänzend wurde dieselbe Analyse getrennt nach Heimdialekt der Teilnehmenden durchgeführt, um allfällige dialektspezifische Verzerrungen zu identifizieren. Die resultierenden Quadrantenverteilungen zeigen zwischen den Heimdialektgruppen keine wesentlichen Unterschiede. Da die Gruppengrössen jedoch stark variieren (*von n=5 beim Wallis bis n=96 bei Zürich*), sind gruppenspezifische Schlussfolgerungen für die meisten Dialekte statistisch nicht belastbar und werden daher nicht weiter vertieft. Einzig für die Gruppe der Zürcher Teilnehmenden ist die Stichprobengrösse ausreichend. Der entsprechende Quadrantenplot ist nahezu identisch mit dem Gesamtplot, was den massgeblichen Einfluss dieser spezifischen Gruppe auf die Gesamtverteilung bestätigt.

4.2.3 Korrelationsanalyse von Sicherheit und Korrektheit

Die Grundannahme des Kalibrierungsmechanismus besagt, dass eine höhere Sicherheit der Teilnehmenden mit einer höheren Korrektheit bei der Dialektidentifikation einhergeht. Zur empirischen Überprüfung dieser Annahme wird der statistische Zusammenhang zwischen der durchschnittlichen Sicherheit und der durchschnittlichen Korrektheit pro Kalibrierungsaufnahme untersucht.

Die Analyse ergibt eine starke positive Pearson-Korrelation von $r = 0.85$. Dies bestätigt, dass Aufnahmen, bei denen sich die Teilnehmenden subjektiv sicher waren, im Durchschnitt auch signifikant häufiger korrekt zugeordnet wurden. Abbildung 20 visualisiert diese lineare Abhängigkeit, wobei die berechnete Regressionsgerade den positiven Trend deutlich macht.

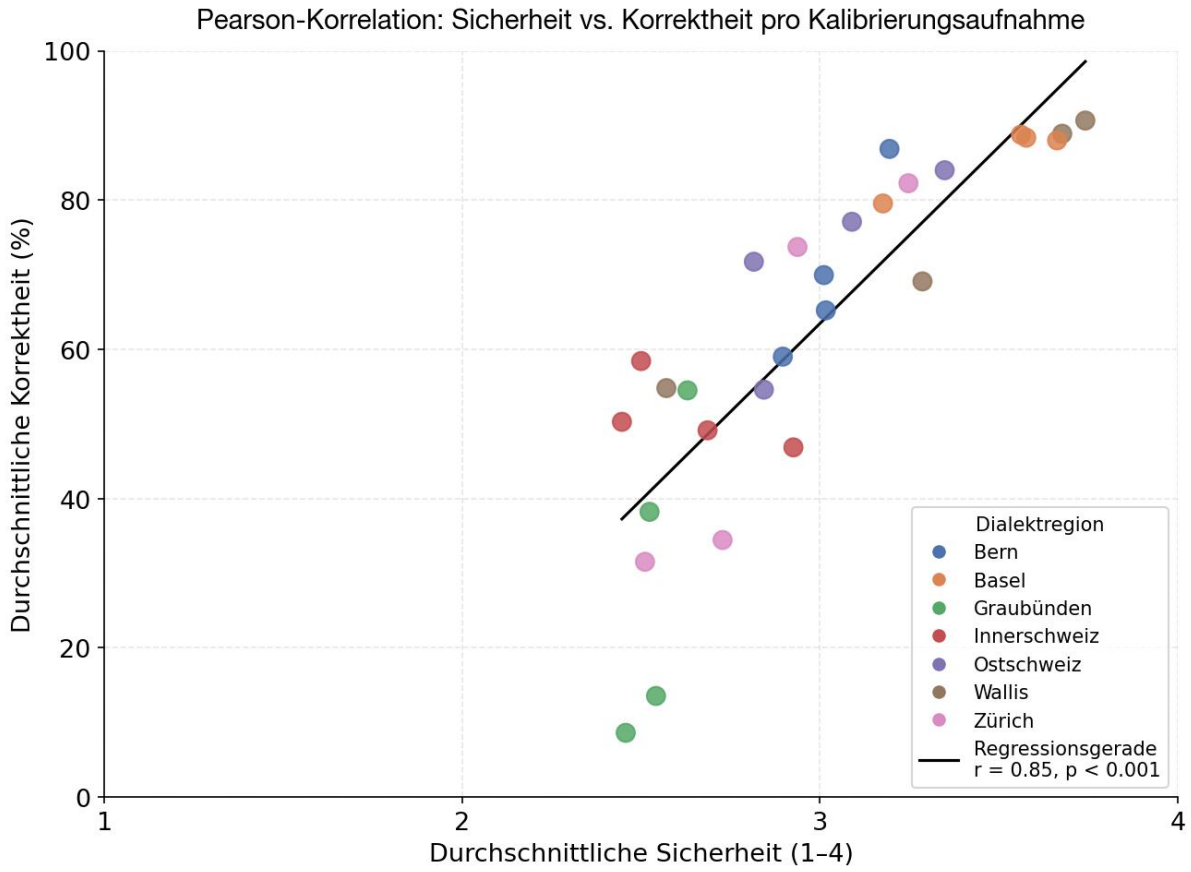

Abbildung 20: Zusammenhang zwischen durchschnittlicher Sicherheit und Korrektheit pro Kalibrierungsaufnahme. Streudiagramm der Kalibrierungsstimuli (farblich differenziert nach Dialektregionen) inklusive linearer Regressionsgeraden. Der Pearson-Korrelationskoeffizient beträgt 0.85.

4.2.4 Eingewöhnungseffekt

Ergänzend stellt sich die Frage, ob die Teilnehmenden zu Beginn der Kalibrierungsphase signifikant weniger korrekte Antworten geben als im weiteren Verlauf, was auf einen initialen Lern- oder Eingewöhnungseffekt hindeuten würde. Zur Überprüfung wird die durchschnittliche Korrektheit in Abhängigkeit der chronologischen Position der Aufnahme (*1 bis 28*) analysiert (*siehe Abbildung 21*).

Auffällig ist ein fast linearer Anstieg der durchschnittlichen Korrektheit über die ersten vier Positionen: ausgehend von einem sehr tiefen Startwert von 13.5 % (*Position 1*), über 38.2 % und 54.5 % bis hin zu 88.4 % bei Position 4. Nach dieser initialen Phase, die bei Position 5 ihren Höhepunkt erreicht (*88.9 %*), zeigt der Verlauf jedoch eine extrem hohe Volatilität ohne kontinuierlichen Aufwärtstrend. Dies wird durch markante Einbrüche, wie etwa der abrupte Abfall auf lediglich 8.6 % bei Position 6, ersichtlich.

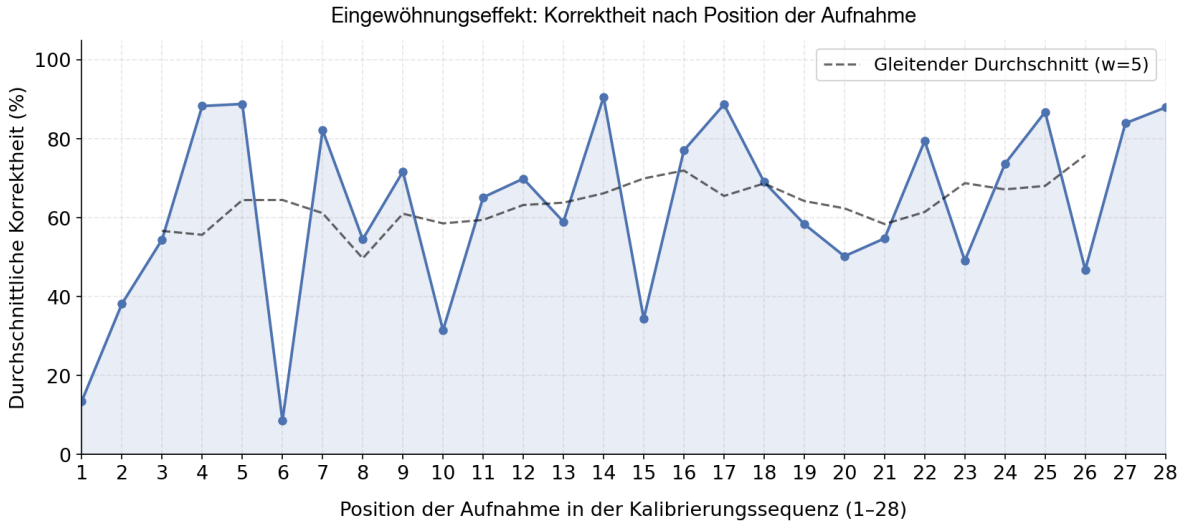

Abbildung 21: Durchschnittliche Korrektheit nach Position der Kalibrierungsaufnahme zur Analyse des Eingewöhnungseffekts. Durchschnittliche Korrektheit bei der Dialektidentifikation (Y-Achse, blaue Linie) in Abhängigkeit zur Position des Stimulus in der Kalibrierungssequenz (X-Achse, 1 bis 28). Die grau gestrichelte Linie zeigt den gleitenden Durchschnitt (Fenstergrösse 5) zur Visualisierung von Trends.

Bei der Interpretation dieser Daten müssen drei massgebliche methodische Einschränkungen berücksichtigt werden, welche die Identifikation eines kausalen Eingewöhnungseffekts limitieren:

- **Retrospektive Korrekturen:** Die eingesetzte Annotationsplattform bot den Teilnehmenden die Möglichkeit, innerhalb der Kalibrierungsphase zurück zu navigieren und bereits abgegebene Antworten nachträglich anzupassen. Ein tatsächlicher Lernprozess während der Aufgabe könnte somit zu retrospektiven Korrekturen geführt haben, was einen potenziellen Effekt in den finalen Daten tendenziell verschleiert.
- **Inhärente Schwierigkeit der Stimuli:** Die stimulusspezifische Schwierigkeit variiert stark. Schwankungen in der Korrektheit werden daher massgeblich durch die linguistischen Merkmale der jeweils gehörten Aufnahme diktiert und nicht zwingend durch den Lernfortschritt der Person.
- **Fehlende Randomisierung:** Die Präsentationsreihenfolge der Kalibrierungsaufnahmen war für alle Teilnehmenden identisch. Durch diese fehlende Randomisierung lässt sich statistisch nicht isolieren, ob die tiefen Werte zu Beginn durch mangelnde Eingewöhnung oder primär durch einen besonders hohen Schwierigkeitsgrad dieser ersten Aufnahmen verursacht wurden.

Aus diesen Gründen lässt sich anhand der vorliegenden Ergebnisse kein belastbarer, systematischer Eingewöhnungseffekt nachweisen. Die beobachteten Varianzen sind in erster Linie auf die stimulusspezifische Schwierigkeit zurückzuführen. Ein isolierter Eingewöhnungseffekt liesse sich methodisch nur durch ein Studiendesign mit vollständig randomisierter Reihenfolge exakt evaluieren.

4.3 Annotationen

Insgesamt wurden im Rahmen der Studie 11’365 Annotationen erhoben. Die Verteilungen der Bewertungen bezüglich der Referenzstimmen sowie die zugrunde liegenden Sätze ist ausgeglichen und weisen keine statistisch signifikanten Unterschiede auf. Dies resultiert direkt aus der randomisierten Präsentationsreihenfolge während der Datenerhebung.

Bei den Dialektregionen zeigt sich hingegen ein anderes Bild (*siehe Abbildung 22*): Da die Mehrheit der Teilnehmenden nicht den gesamten verfügbaren Annotationsdatensatz bewertet hat, griff der algorithmische Steuerungsmechanismus (*siehe Kapitel 3.1.3*). Basierend auf den individuellen Kalibrierungsergebnissen wurden den Teilnehmenden bevorzugt jene Stimuli präsentiert, die sie voraussichtlich präziser bewerten konnten. Entsprechend spiegelt die Verteilung der gesammelten Annotationen nach Dialekten weitgehend die allgemeinen Kalibrierungsscores wider (*siehe Abbildung 17*).

Ein Grossteil der Annotationen stammt von Personen, die den bewerteten Dialekt nicht als eigenen Heimdialekt sprechen (*Fremdbewertung*). Eine markante Ausnahme bildet hierbei die Dialektregion Zürich, bei der die Eigenbewertungen dominieren. Dieser Effekt ist die direkte methodische Konsequenz der in Kapitel 4.1 dokumentierten demografischen Überrepräsentation der Zürcher Stichprobe.

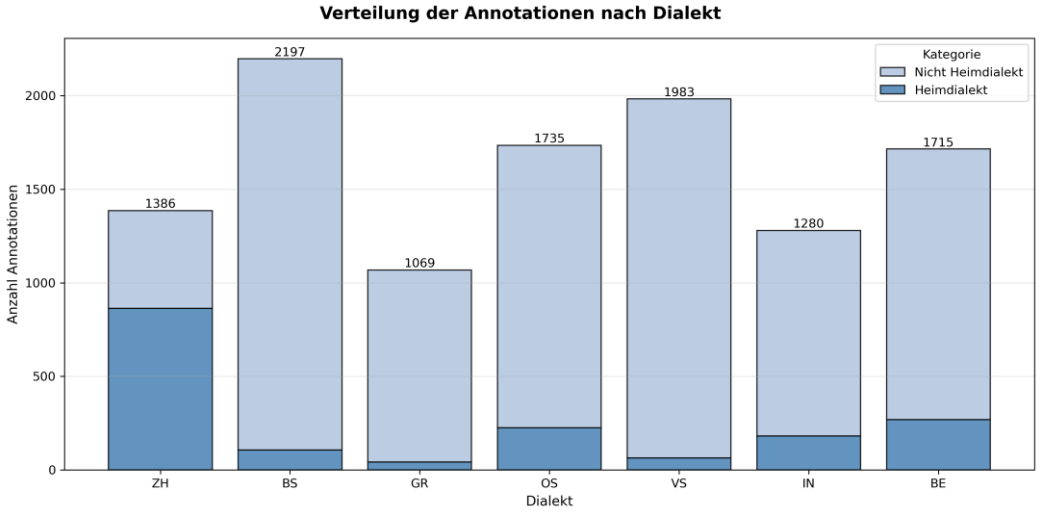

Abbildung 22: Verteilung der Annotationen nach Dialekt. Gesamtzahl der abgegebenen Annotationen (Y-Achse) für die sieben Dialektregionen (X-Achse). Die gestapelten Säulen unterscheiden zwischen Übereinstimmung mit dem Heimdialekt (dunkelblau) und Abweichung (hellblau). Die exakten Werte stehen über den Säulen.

Die Analyse der erhobenen Sicherheitswerte – also der subjektiven Gewissheit der Teilnehmenden während der eigentlichen Annotationsphase – liefert weitere wichtige methodische Erkenntnisse (*siehe Abbildung 23*). Bezüglich der Referenzstimmen und Sätze zeigen sich auch hier keine statistisch relevanten Abweichungen, was die homogene Verteilung der Stimuli bestätigt.

Bei den Dialektregionen tritt jedoch erneut der strukturelle Einfluss der Zürcher Stichprobe hervor: Die Sicherheitswerte für den Zürcher Dialekt liegen im Durchschnitt signifikant höher als bei allen anderen Dialektregionen. Dies lässt sich auf den inhärenten Vertrautheitseffekt (*In-Group-Effekt*) der Zürcher Mehrheit zurückführen.

Auffällig ist, dass die Dialektregion Innerschweiz (*IN*) im direkten Vergleich mit allen anderen Regionen deutlich tiefere Sicherheitswerte verzeichnet. Für diese systematische Unsicherheit lassen sich zwei potenzielle Erklärungsansätze anführen:

- **Phonetische Varianz und Fragmentierung:** Der Innerschweizer Dialektraum ist linguistisch stark fragmentiert (*z. B. Luzerner, Urner oder Unterwaldner Dialekt*). Diese internen Varianzen weisen teilweise sehr unterschiedliche Lautsysteme auf. Das Fehlen des einen, homogenen Innerschweizer Prototyps in den Audiobeispielen könnte die Teilnehmenden bei der Zuordnung systematisch verunsichert haben.

- **Wahrnehmungsasymmetrie:** Wie die Auswertung der Selbsteinschätzung in Kapitel 4.1 bereits gezeigt hat, stuften die Teilnehmenden ihre eigene Erkennungskompetenz für den Innerschweizer Dialekt schon vorab am geringsten ein. Die tiefen Sicherheitswerte in der Annotationsphase dürften diese subjektive Barriere widerspiegeln. Es ist davon auszugehen, dass die Teilnehmenden kein gefestigtes inneres Referenzsystem für diesen Dialekt besassen, was in der Folge zu einer dauerhaft höheren Verunsicherung beim Hören der Stimuli führte.

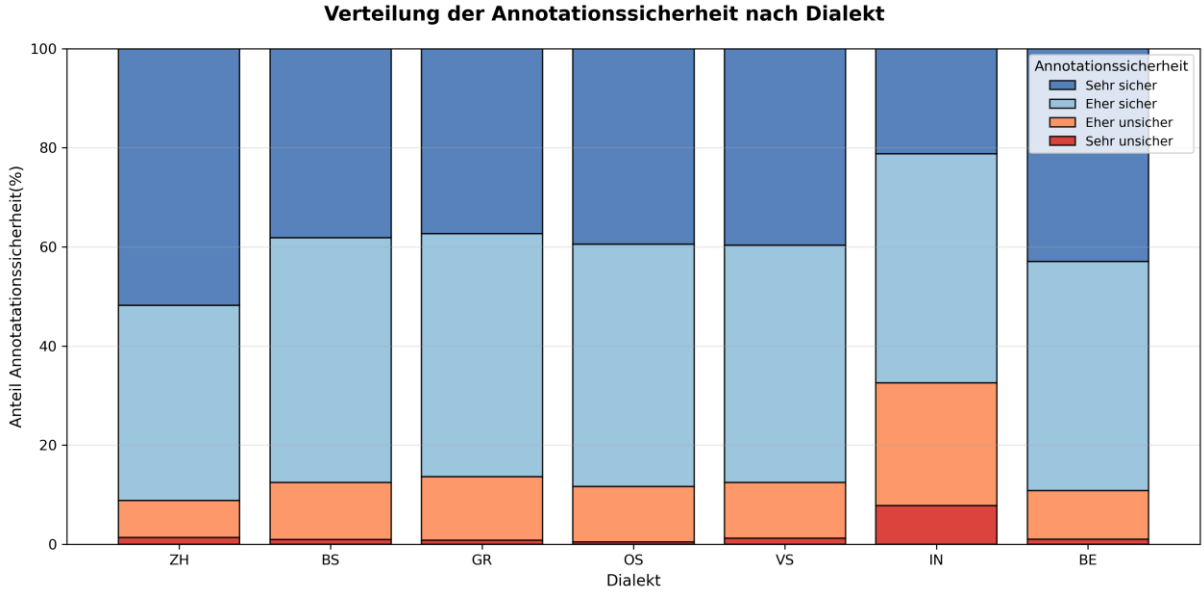

Abbildung 23: Verteilung der Annotationssicherheit nach Dialekt. Prozentualer Anteil der vier Sicherheitsstufen (Y-Achse) aufgeteilt nach evaluiertem Dialekt (X-Achse). Die Farben der 100-%-gestapelten Säulen codieren die Annotationssicherheit von «Sehr sicher» (dunkelblau) bis «Sehr unsicher» (rot).

5 Auswertungsmethoden

Zur Auswertung der gesammelten Annotationen werden zwei statistische Methoden evaluiert. Ziel ist es, aus den individuellen Bewertungen der Teilnehmenden eine robuste, aggregierte Aussage über die wahrgenommene Authentizität der Stimuli zu gewinnen und die verschiedenen Einflussfaktoren statistisch zu quantifizieren.

Ein initial geprüfter Ansatz basierend auf einem Mehrheitsentscheid (*Majority Vote*) wurde für die finale Auswertung verworfen, da die IRR weit unter einem akzeptablen Schwellenwert lag. Als massgebliches Auswertungsverfahren kommt stattdessen die Bayessche ordinale logistische Regression zum Einsatz.

5.1 Majority Vote

Jeder Stimulus i wird von n_i Teilnehmenden auf einer ordinalen Skala von 1 bis 4 bewertet. Zur Ermittlung einer finalen Stufe $R_i \in \{1, 2, 3, 4\}$ pro Stimulus bedarf es einer Aggregationsmethode. Da die Teilnehmenden unterschiedliche Dialektkenntnisse sowie variierende Sicherheiten bei den einzelnen Bewertungen aufweisen, stellt sich die methodische Frage, ob alle Bewertungen gleich gewichtet werden sollen oder ob eine Gewichtung nach Kompetenz und Sicherheit zielführend ist. Es werden drei Varianten berechnet und verglichen, die methodisch aufeinander aufbauen:

Naive Majority Vote (*Baseline*): Die am häufigsten gewählte Stufe wird zugewiesen:

$$R_i = \text{mode}(r_{i1}, r_{i2}, \dots, r_{in_i})$$

Gewichtung nach Annotationssicherheit: Jede Bewertung r_{ij} wird mit dem stimulusspezifischen Sicherheitsmass $c_{ij} \in \{1, 2, 3, 4\}$ gewichtet. Pro Stufe k wird die gewichtete Häufigkeit berechnet:

$$w_{ij}^{\text{ann}} = \frac{c_{ij}}{4}$$

$$S_k = \sum_{j:r_{ij}=k} w_{ij}^{\text{ann}}$$

$$R_i = \arg\max_k S_k$$

Gemischte Gewichtung: Diese Variante kombiniert die stimulusspezifische Sicherheit mit dem Kalibrierungsscore $s_j^d \in$ [-1, 1] der Person j für den Dialekt d des jeweiligen Stimulus. Der Kalibrierungsscore wird dazu auf $[0, 1]$ normalisiert. Die isolierte Verwendung der Kalibrierungsresultate entfällt, da diese keine Aussage über die Sicherheit bei einem spezifischen Stimulus erlaubt. Das Gewichtungsverhältnis wird durch die Parameter α_{ann} und α_{cal} (*mit* $\alpha_{ann} + \alpha_{cal} = 1$) gesteuert:

$$w_{ij}^{\text{mix}} = \alpha_{\text{ann}} \cdot \frac{c_{ij}}{4} + \alpha_{\text{cal}} \frac{s_j^d + 1}{2}$$

$$S_k = \sum_{j:r_{ij}=k} w_{ij}^{\text{mix}}$$

$$R_i = \arg\max_k S_k$$

Fehlt für eine teilnehmende Person der Kalibrierungsscore, gilt: $w_{ij}^{\text{mix}} = w_{ij}^{\text{ann}}$.

Da das optimale Verhältnis der beiden Parameter α_{cal} und α_{ann} nicht *a priori* bekannt ist, erfolgt eine Grid-Search-Optimierung über 13 diskrete Kombinationen mit $\alpha_{\text{cal}} \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$, wobei die mathematische Bedingung $\alpha_{\text{ann}} = 1 - \alpha_{\text{cal}}$ gilt. Als Optimierungskriterium dient die IRR in Form von Krippendorffs Alpha mit Intervalldistanz. Diejenige Parameterkombination, welche die höchste Übereinstimmung erzielt, wird schliesslich für die finale Auswertung herangezogen.

Bei einem Gleichstand (*Tie*) in der finalen Stufenwahl entscheidet in allen drei Varianten der gerundete Median der jeweiligen Bewertungen als Tie-Breaker.

5.1.1 IRR

Die IRR wird für alle drei Auswertungsvarianten berechnet. Als Basis dient die in Kapitel 2.4.1 definierte Standardformulierung von Krippendorffs Alpha. Für den naiven Mehrheitsentscheid wird die quadrierte, normalisierte Intervalldistanz verwendet.

Abgewandelte Form mit sicherheitsgewichteter Distanz: Um der variierenden Sicherheit der Teilnehmenden Rechnung zu tragen, wird eine abgewandelte Distanzmetrik eingeführt. Dabei wird die quadrierte, normalisierte Intervalldistanz mit der durchschnittlichen Sicherheit der beiden an der Berechnung beteiligten Bewertenden multipliziert:

$$\delta^2(c,k) = \left(\frac{|c-k|}{\max(r) - \min(r)}\right)^2 \cdot \frac{x_c + x_k}{2 \cdot x_{\max}}$$

In dieser Gleichung stehen c und k für die Bewertung (Klasse) der zwei Teilnehmenden. Die Variablen x_c und x_k repräsentieren die dazugehörigen angegebenen Annotationssicherheiten auf einer Skala von 1 bis 4. Der Nennerterm $x_{\max}$ definiert den maximalen theoretischen Sicherheitswert (*hier 4*) über alle Bewertungen, um den Faktor auf den Bereich $[0,1]$ zu skalieren. Diese Modifikation der Distanzfunktion führt dazu, dass eine Abweichung (*Disagreement*) das resultierende Alpha zwischen zwei Teilnehmenden, die sich ihrer Bewertung sicher waren, stärker penalisiert als eine Abweichung zwischen zwei unsicheren Teilnehmenden. Dies ist methodisch sinnvoll, da Bewertungen mit tiefer Sicherheit einen geringeren Informationsgehalt aufweisen und folglich die Zuverlässigkeit der Gesamtmetrik weniger stark belasten sollte.

Für die Grid-Search-Optimierung werden die Sicherheitsvariablen durch das gemischte Gewicht w_{ij}^{mix} ersetzt, wobei der Nenner die maximale gemischte Gewichtung von 1 abbildet:

$$\delta^2(c,k) = \left(\frac{|c-k|}{\max(r) - \min(r)}\right)^2 \cdot \frac{w_c^{\text{mix}} + w_k^{\text{mix}}}{2}$$

Die Optimierung ermittelt jene Parameterkombination ($\alpha_{\text{cal}}, \alpha_{\text{ann}}$), welche den finalen Alpha-Wert maximiert und somit die Gewichtung identifiziert, bei der die sicheren und dialektkompetenten Teilnehmenden die höchste Konsistenz aufweisen.

Wie in Tabelle 2 ersichtlich, liegen die berechneten Alpha-Werte für alle drei untersuchten Varianten in der aktuellen Studie zwischen 0.19 und 0.21. Damit verfehlen sie den akzeptablen Schwellwert von $\alpha \geq 0.67$ deutlich. Die beobachteten Unterschiede zwischen der naiven Baseline und den gewichteten Varianten sind zwar konsistent in ihrer Richtung, jedoch in ihrer absoluten Grössenordnung marginal.

Tabelle 2: Krippendorffs Alpha pro Majority Vote Variante. Gegenüberstellung von drei Evaluationsansätzen (Naive Baseline, gewichtet nach Sicherheit, 50/50 gemischt) und den daraus resultierenden Reliabilitätswerten (Krippendorffs Alpha).

Majority Vote Variante	**Krippendorf's α**
Naive (Baseline)	0.1989
Gewichtet Annotations-Sicherheit	0.2094
Gewichtet Gemischt (50/50)	0.2084

Die durchgehend tiefen Werte deuten darauf hin, dass keine substanzielle Übereinstimmung zwischen den Teilnehmenden bezüglich der wahrgenommenen Qualität der Stimuli besteht. Aufgrund dieser tiefen IRR kann der Majority Vote nicht als zuverlässiges Aggregationsmass für die finale Stufenzuteilung betrachtet werden. Die zugeteilten Stufen sind unter diesen Voraussetzungen nicht als stabile, reproduzierbare Aussagen über die einzelnen Stimuli zu interpretieren.

5.2 Spezifikation der bayesschen Modelle

Auf Basis der theoretischen Grundlagen (*siehe Kapitel 2.4.2*) werden im Folgenden die spezifischen Modelle zur Auswertung der Annotationsstudie definiert. Ziel ist die Quantifizierung isolierten und kombinierten Einflüsse verschiedener Merkmale. Diese lassen sich in stimulusspezifische Faktoren (*Dialekt, Referenzstimme und Satz*) sowie personenspezifische Faktoren (*demografische Eigenschaften, Annotationssicherheit und Kalibrierungsdaten*) unterteilen. Die Evaluation und der Modellvergleich erfolgen durchgehend anhand der in Kapitel 2.4.3 beschriebenen Validierungsmetriken.

Zur methodischen Isolation dieser Effekte wurde ein inkrementeller Modellaufbau gewählt (*siehe Tabelle 3*). Ausgehend von einem Basismodell (*M1*) werden schrittweise systematische Einflussfaktoren als Prädiktoren hinzugefügt.

Tabelle 3: Übersicht der bayesschen ordinalen logistischen Regressionsmodelle mit jeweiliger Basis und Erweiterung. Auflistung der Modellarchitekturen (M1 bis M9.2). Die Spalten zeigen die verwendete Abkürzung, das jeweilige Basismodell sowie die inhaltliche Beschreibung der sequenziell hinzugefügten Variablen.

Modell	Abkürzung	Basis	Beschreibung
M1	Base	–	Basismodell mit Modellierung der personenspezifischen Bias
M2.1	Base+Dialekt	M1	Hinzunahme von Dialekteffekten
M2.2	Base+Stimme	M1	Hinzunahme von Referenzstimmeffekten
M2.3	Base+Satz	M1	Hinzunahme von Satzeffekten
M3	Base+Stimulus	M1	Hinzunahme von Dialekt-, Referenzstimm- und Satzeffekten
M4	Full	M3	Hinzunahme der Annotationssicherheit
M5.1	Base+Heim	M1	Hinzunahme des Heimdialekts der Teilnehmenden
M5.2	Full+Heim	M4	Hinzunahme des Heimdialekts der Teilnehmenden
M6.1	Base+Selbst	M2.1	Hinzunahme der Selbsteinschätzung (Dialektidentifikation)
M6.2	Full+Selbst	M4	Hinzunahme der Selbsteinschätzung (Dialektidentifikation)
M7.1	Base+Kalib	M2.1	Hinzunahme der Kalibrierungsscores
M7.2	Full+Kalib	M4	Hinzunahme der Kalibrierungsscores
M8.1	Base+Alter	M1	Hinzunahme der demografischen Altersgruppen
M8.2	Full+Alter	M4	Hinzunahme der demografischen Altersgruppen
M9.1	Base+Expo	M2.1	Hinzunahme der geografischen Dialektexposition
M9.2	Full+Expo	M4	Hinzunahme der geografischen Dialektexposition

Die Modelle lassen sich funktional in folgende Gruppen einteilen:

- **Aufbau des Basismodells (*M1 bis M3*):** Ausgehend von einem Basismodell für den Bias der Teilnehmenden (*M1*) werden schrittweise stimulusspezifische Effekte (*Dialekt, Referenzstimme und Sätze*) integriert (*M2 bis M3*).
- **Einführung des Kontrollmodells (*M4*):** Darauf aufbauend wird die subjektive Annotationssicherheit (*M4*) als Kontrollparameter eingeführt. Dieses Modell bündelt alle stimulusspezifischen Effekte und dient im weiteren Verlauf als Kontrollmodell.
- **Zweistufige Prüfung individueller Parameter (*M5 bis M9*):** In den nachfolgenden Modellgruppen werden gezielt demografische und personenspezifische Parameter untersucht (*Heimdialekt, Dialektkenntnisse, Kalibrierungsleistung, Alter, geografische Exposition*). Um die Robustheit dieser Effekte zu prüfen, wird jeder Parameter zweistufig modelliert:
  - **Isolierte Schätzung:** Zunächst wird der Effekt in einem reduzierten Basismodell (*M1 oder dem dialektspezifischen Modell M2.1*) geschätzt.
  - **Kontrollierte Verifikation:** Anschliessend wird der Prädiktor in das Kontrollmodell M4 integriert, um zu prüfen, ob der Erklärungsbeitrag auch bei gleichzeitiger Kontrolle aller anderen Variablen bestehen bleibt.

Dieser systematische Aufbau ermöglicht es, den Erklärungsbeitrag jedes einzelnen Prädiktors auf die wahrgenommene Qualität der Stimuli isoliert und statistisch belastbar zu bewerten.

5.2.1 Modellgruppe M1: Personenspezifischer Bias

Dieses Modell dient als Basis für die weiteren Analysen. Es fungiert als Referenz für den quantitativen Vergleich der Modellgüte bei der schrittweisen Integration zusätzlicher Einflussfaktoren in den Folgemodellen.

Die Struktur des linearen Prädiktors μ_i für eine spezifische Bewertung $i \in \{1, \ldots, N\}$ ist hierarchisch wie folgt definiert:

$$\begin{aligned} \sigma_{\text{teilnehmende}} &\sim \text{Exponential}(1) \\ \beta_{\text{teilnehmende}[p]} &\sim \mathcal{N}(0, \sigma_{\text{teilnehmende}}) \\ \tilde{\mu}_i &= \beta_{\text{teilnehmende}[p[i]]} \\ \mu_i &= \tilde{\mu}_i - \frac{1}{N}\sum_{j=1}^{N} \tilde{\mu}_j \end{aligned}$$

Der latente Erwartungswert für eine Bewertung ergibt sich in diesem Modell ausschliesslich aus dem personenspezifischen Bias. Die Parameter $\beta_{\text{teilnehmende}[p]}$ der P Teilnehmenden werden aus einer gemeinsamen Normalverteilung gezogen, deren Streuung $\sigma_{\text{teilnehmende}}$ empirisch über einen Exponential-Prior geschätzt wird. Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ verknüpft jede Bewertung i mit dem Effekt der entsprechenden Person $p[i]$.

Die Zentrierung erfolgt anschliessend auf der Ebene der N Gesamtbewertungen. Die Subtraktion des Mittelwerts aller unzentrierten Prädiktoren $\tilde{\mu}$ stellt mathematisch sicher, dass die finalen Werte μ_i als relative Abweichungen interpretiert werden können. Dies ist essenziell, um die Identifizierbarkeit des Modells im Kontext der ordinalen Schwellenwerte (*Cutpoints*) zu gewährleisten. Die Teilnehmenden werden hierbei als zufällige Effekte (*Random Effects*) modelliert, um individuelle Antworttendenzen (*z. B. generelle Strenge oder Milde bei der Bewertung*) statistisch abzufangen. Die genaue Ausprägung dieses Bias ist nicht von primärem Interesse in dieser Arbeit und wird nicht isoliert interpretiert. Seine Modellierung ist jedoch essenziell: Sie kontrolliert die personenspezifische Varianz und stellt sicher, dass in den Folgemodellen sowohl stimulusspezifischen als weitere personenspezifischen Effekte unverfälscht gemessen werden.

5.2.2 Modellgruppen M2, M3: Stimulusspezifische Effekte

Die folgenden Modelle erweitern das Basismodell schrittweise um zusätzliche hierarchische Einflussfaktoren auf Dialekt-, Referenzstimme- und Satzebene. Ziel ist es, deren Erklärungsbeitrag auf die Bewertungen zu quantifizieren und die Modellgüte systematisch zu vergleichen. Alle Modelle übernehmen den hierarchischen Aufbau zur Modellierung des personenspezifischen Bias aus Modell M1 (*Base*). Ergänzend zu diesem Bias werden Dialekt-, Referenzstimmen- und Satzeffekte als strukturelle Einflussfaktoren modelliert.

Modell M2.1 (*Base+Dialekt*) erweitert die Struktur um Dialekteffekte. Für eine spezifische Bewertung $i \in \{1, \ldots, N\}$ und die Menge aller Dialekte $d \in \{1, \ldots, D\}$ wird das Modell wie folgt definiert:

$$\begin{aligned} \sigma_{\text{dialekt}} &\sim \text{Exponential}(1) \\ z_{\text{dialekt}[d]} &\sim \mathcal{N}(0,1) \\ \tilde{\beta}_{\text{dialekt}[d]} &= z_{\text{dialekt}[d]} \cdot \sigma_{\text{dialekt}} \\ \beta_{\text{dialekt}[d]} &= \tilde{\beta}_{\text{dialekt}[d]} - \frac{1}{D}\sum_{j=1}^{D} \tilde{\beta}_{\text{dialekt}[j]} \\ \tilde{\mu}_i &= \beta_{\text{dialekt}[d[i]]} + \beta_{\text{teilnehmende}[p[i]]} \end{aligned}$$

Zur Modellierung der Dialekteffekte wird ein hierarchischer, nicht zentrierter Ansatz (*Non-Centered Parameterization*) verwendet. Dabei wird zunächst eine standardnormalverteilte

Variable $z_{\text{dialekt}[d]} \sim \mathcal{N}(0,1)$ eingeführt, welche durch den skalierenden Hyperparameter σ_{dialekt} in den effektiven, aber noch nicht zentrierten Regressionskoeffizienten $\tilde{\beta}_{\text{dialekt}[d]}$ überführt wird. Diese Form der Parametrisierung verbessert die geometrischen Eigenschaften des Posteriorraums und führt in der Regel zu einer effizienteren MCMC-Inferenz. Die Dialektrohparameter werden anschliessend auf der Ebene der D Dialekte zentriert, sodass ihre Summe null ergibt.

Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ für eine spezifische Bewertung i setzt sich additiv aus dem zentrierten Dialekteffekt $d[i]$ und dem zugehörigen personenspezifischen Bias $p[i]$ zusammen. Wie bereits im Basismodell (*M1*) definiert, erfolgt für alle Modelle abschliessend die finale globale Zentrierung dieses Prädiktors ($\mu_i = \tilde{\mu}_i - \frac{1}{N}\sum_{j=1}^{N} \tilde{\mu}_j$) über alle N Gesamtbewertungen, um die Identifizierbarkeit sicherzustellen.

Modell M2.2 (*Base+Stimme*) folgt der exakt gleichen methodischen Struktur wie M2.1, ersetzt jedoch die Dialektebene durch Referenzstimmeffekte β_{stimme} für die Referenzstimmen $s \in \{1, \ldots, S\}$:

$$\tilde{\mu}_i = \beta_{\text{stimme}[s[i]]} + \beta_{\text{teilnehmende}[p[i]]}$$

Modell M2.3 (*Base+Satz*) folgt ebenfalls dieser Struktur und ersetzt die Dialektebene durch Satzeffekte β_{satz} für die Sätze $u \in \{1, \ldots, U\}$:

$$\tilde{\mu}_i = \beta_{\text{satz}[u[i]]} + \beta_{\text{teilnehmende}[p[i]]}$$

Modell M3 (*Base+Stimulus*) vereint alle drei stimulusspezifischen Effekte gleichzeitig. Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ setzt sich dabei additiv aus den Dialekt-, Referenzstimmen-, Satzeffekten sowie dem personenspezifischen Bias zusammen, um deren kombinierten Einfluss auf die Bewertung zu quantifizieren:

$$\tilde{\mu}_i = \beta_{\text{dialekt}[d[i]]} + \beta_{\text{stimme}[s[i]]} + \beta_{\text{satz}[u[i]]} + \beta_{\text{teilnehmende}[p[i]]}$$

5.2.3 Modellgruppe M4: Annotationssicherheit

Das Modell M4 (*Full*) erweitert das vereinte stimulusspezifische Modell M3 (*Base+Stimulus*) um die von den Teilnehmenden pro Annotation gegebene Sicherheit. Diese wird als personenspezifische Sensitivität gegenüber Unsicherheiten modelliert, indem ein individueller Steigungsparameter (*Slope*) pro teilnehmender Person eingeführt wird. Formal wird zunächst die Variabilität dieser Sicherheitsslopes hierarchisch modelliert:

$$\sigma_{\text{sicherheit}} \sim \text{Exponential}(1)$$
$$\beta_{\text{sicherheit}[p]} \sim \mathcal{N}(0, \sigma_{\text{sicherheit}})$$

Um die numerische Stabilität der Modellierung zu gewährleisten und die Effekte interpretierbar zu machen, werden die diskreten, rohen Sicherheitswerte x_i vorab über alle Datenpunkte hinweg standardisiert. Der standardisierte Sicherheitswert a_i für eine spezifische Bewertung i

berechnet sich durch die Subtraktion des empirischen Mittelwerts $\bar{x}$ und die Division durch die Standardabweichung $\mathrm{sd}(x)$:

$$a_i = \frac{x_\mathrm{i} - \bar{x}}{\mathrm{sd}(x)}$$

Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ setzt sich anschliessend aus den Dialekt-, Referenzstimmen-, Satzeffekten, dem personenspezifischen Bias sowie dem neuen interaktiven Sicherheitseffekt zusammen. Letzterer ergibt sich aus der Multiplikation des personenspezifischen Slopes $\beta_{\mathrm{sicherheit}[p[i]]}$ mit der standardisierten Sicherheit a_i der jeweiligen Bewertung:

$$\tilde{\mu}_i = \beta_{\mathrm{dialekt}[d[i]]} + \beta_{\mathrm{stimme}[s[i]]} + \beta_{\mathrm{satz}[u[i]]} + \beta_{\mathrm{teilnehmende}[p[i]]} + (\beta_{\mathrm{sicherheit}[p[i]]} \cdot a_i)$$

5.2.4 Modellgruppe M5: Heimdialekt

Die Modellgruppe M5 untersucht den Einfluss des Heimdialekts auf das Bewertungsverhalten. Konkret wird dabei modelliert, ob die Übereinstimmung zwischen dem im Stimulus gesprochenen Dialekt und der eigenen Dialektregion der teilnehmenden Person einen Einfluss auf das Bewertungsverhalten hat.

Das Modell M5.1 (*Base+Heim*) isoliert diesen Effekt und erweitert, analog zu den Modellen der M2-Gruppe, das Basismodell M1 um die Angabe des Heimdialekts. Für eine spezifische Bewertung i und den binären Indikator $h \in \{0,1\}$ (*Übereinstimmung ja/nein*) wird der unzentrierte lineare Prädiktor wie folgt definiert:

$$\tilde{\mu}_i = \beta_{\mathrm{heimdialekt}[h[i]]} + \beta_{\mathrm{teilnehmende}[p[i]]}$$

Das Modell M5.2 (*Full+Heim*) verwendet die umfassende Struktur aus dem kombinierten Modell M4. Ziel ist es, den Einfluss des Heimdialekts unter Kontrolle aller bisherigen stimulus- und personenspezifischen Faktoren (*inklusive der Annotationssicherheit*) zu evaluieren:

$$\tilde{\mu}_i = \beta_{\mathrm{dialekt}[d[i]]} + \beta_{\mathrm{stimme}[s[i]]} + \beta_{\mathrm{satz}[u[i]]} + \beta_{\mathrm{teilnehmende}[p[i]]} + (\beta_{\mathrm{sicherheit}[p[i]]} \cdot a_i) + \beta_{\mathrm{heimdialekt}[h[i]]}$$

Auch für den Heimdialekteffekt kommt die in Kapitel 5.2.2 eingeführte hierarchische, nicht zentrierte Parametrisierung zum Einsatz. Wie bei allen Modellen wird der Prädiktor $\tilde{\mu}_i$ abschliessend über alle N Bewertungen zentriert, um die Identifizierbarkeit des Gesamtmodells zu gewährleisten.

5.2.5 Modellgruppe M6: Selbsteinschätzung

Die Modellgruppe M6 untersucht den Einfluss der subjektiven Selbsteinschätzung (*bezüglich der eigenen Dialektidentifikation*) auf das Bewertungsverhalten.

Die im Fragebogen erhobenen rohen Antworten zur Selbsteinschätzung werden vorab auf das Intervall $[0,1]$ skaliert, wobei die Angabe «nicht vertraut» zur Vereinfachung des Modells den Wert 0 erhält. Analog zur Annotationssicherheit wird dieser skalierte Erkennungswert $x_{erkennung}$ anschliessend über alle Datenpunkte hinweg standardisiert, um die MCMC-Inferenz zu stabilisieren. Der standardisierte Wert r_i für eine spezifische Bewertung i berechnet sich somit als:

$$r_i = \frac{x_{erkennung,i} - \bar{x}_{erkennung}}{\mathrm{sd}(x_{erkennung})}$$

Im Modell M6.1 (*Base+Selbst*) wird das auf Dialekteffekte beschränkte M2.1 um diesen neuen Prädiktor ergänzt. Aufgrund der geringen Anzahl an spezifischen Datenpunkten pro teilnehmender Person wird die Selbsteinschätzung hierbei – im Gegensatz zu den vorherigen Parametern – nicht hierarchisch modelliert, sondern als globaler Haupteffekt (*Fixed Effect*) über alle Teilnehmenden hinweg geschätzt. Dem Parameter wird ein schwach informativer Normalverteilungs-Prior zugewiesen:

$$\beta_{\mathrm{erkennung}} \sim \mathcal{N}(0,1)$$

Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ setzt sich in Modell M6.1 additiv aus den Dialekteffekten, dem personenspezifischen Bias sowie der gewichteten, standardisierten Selbsteinschätzung zusammen:

$$\tilde{\mu}_i = \beta_{dialekt[d[i]]} + \beta_{teilnehmende[p[i]]} + (\beta_{erkennung} \cdot r_i)$$

Das Modell M6.2 (*Full+Selbst*) integriert diesen neuen Prädiktor in die Modellstruktur von Modell M4. Dadurch kann der isolierte Effekt der Selbsteinschätzung verifiziert werden, während gleichzeitig alle stimulusspezifischen Effekte (*Dialekt, Referenzstimme, Satz*) sowie die individuelle Annotationssicherheit (a_i) kontrolliert werden:

$$\tilde{\mu}_i = \beta_{\mathrm{dialekt}[d[i]]} + \beta_{\mathrm{stimme}[s[i]]} + \beta_{\mathrm{satz}[u[i]]} + \beta_{\mathrm{teilnehmende}[p[i]]} + (\beta_{\mathrm{sicherheit}[p[i]]} \cdot a_i) + (\beta_{erkennung} \cdot r_i)$$

Wie bei den vorangegangenen Modellen erfolgt die finale Transformation zum auswertbaren Prädiktor μ_i durch eine globale Zentrierung über alle N Bewertungen.

5.2.6 Modellgruppe M7: Kalibrierungsscore

Die Modellgruppe M7 untersucht den Einfluss der tatsächlichen, in der Kalibrierungsphase empirisch gemessenen Erkennungsleistung der Teilnehmenden auf deren anschliessendes Bewertungsverhalten.

Der in Kapitel 3.1.3 definierte Dialektscore C_d einer Person für einen spezifischen Dialekt wird hier als personenspezifischer Kalibrierungsscore $c_{p,d}$ formalisiert. Um Skaleneffekt zu eliminieren und die numerische Stabilität des MCMC-Samplers zu gewährleisten, werden diese Werte über alle Beobachtungen hinweg standardisiert.

Der standardisierte Kalibrierungsscore k_i für eine spezifische Bewertung i, welche von der teilnehmenden Person $p[i]$ für einen Stimulus des Dialekts $d[i]$ abgegeben wurde, berechnet sich somit als:

$$k_i = \frac{c_{p[i],d[i]} - \bar{c}}{\mathrm{sd}(c)}$$

Im Modell M7.1 (*Base+Kalib*) wird dieser Prädiktor in das rein dialektspezifische Modell M2.1 integriert. Da pro Kombination aus teilnehmender Person und Dialekt nur eine sehr begrenzte Anzahl an Kalibrierungsdaten vorliegt, wird der Effekt nicht hierarchisch pro Individuum strukturiert, sondern als globaler Haupteffekt (*Fixed Effect*) über den gesamten Datensatz hinweg

geschätzt. Dem Regressionskoeffizienten wird ein schwach informativer Normalverteilungs-Prior zugewiesen:

$$\beta_{kalibrierung} \sim \mathcal{N}(0,1)$$

Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ für eine Bewertung i setzt sich in Modell M7.1 additiv aus dem Dialekteffekt, dem personenspezifischen Bias sowie der mit dem globalen Koeffizienten gewichteten, standardisierten Kalibrierungsleistung zusammen:

$$\tilde{\mu}_i = \beta_{dialekt[d[i]]} + \beta_{teilnehmende[p[i]]} + (\beta_{kalibrierung} \cdot k_i)$$

Das Modell M7.2 (*Full+Kalib*) führt diesen Prädiktor schliesslich in die Modellstruktur von Modell M4 ein. Dies erlaubt es, den isolierten Beitrag der tatsächlichen Kalibrierungsleistung zu analysieren, während gleichzeitig alle stimulusspezifischen Störfaktoren (Dialekt, Referenzstimme, Satz) sowie die subjektive Annotationssicherheit (a_i) statistisch kontrolliert werden:

$$\tilde{\mu}_i = \beta_{\text{dialekt}[d[i]]} + \beta_{\text{stimme}[s[i]]} + \beta_{\text{satz}[u[i]]} + \beta_{\text{teilnehmende}[p[i]]} + (\beta_{\text{sicherheit}[p[i]]} \cdot a_i) + (\beta_{kalibrierung} \cdot k_i)$$

Die Überführung in den finalen Prädiktor μ_i erfolgt auch in dieser Modellgruppe durch die globale Zentrierung über alle N Gesamtbewertungen, um die mathematische Identfizierbarkeit zu wahren.

5.2.7 Modellgruppe M8: Altersgruppen

Die Modellgruppe M8 untersucht den Einfluss von Altersgruppen auf das Bewertungsverhalten. Im Unterschied zu vorherigen Parametern handelt es sich hierbei um ein Merkmal, welches vollständig der teilnehmenden Person zugewiesen ist. Aus diesem Grund muss der Effekt hierarchisch als Teil des personenspezifischen Bias modelliert werden, da ansonsten die beiden Effekte beim Sampling um die gleiche Varianz konkurrenzieren würden.

Für eine Altersgruppe $k \in \{1, \ldots, K\}$ und eine teilnehmende Person $p \in \{1, \ldots, P\}$ werden die Zusammenhänge wie folgt definiert:

$$\begin{aligned}
\sigma_{\text{alter}} &\sim \text{Exponential}(1) \\
z_{\text{alter}[k]} &\sim \mathcal{N}(0,1) \\
\tilde{\beta}_{\text{alter}[k]} &= z_{\text{alter}[k]} \cdot \sigma_{\text{alter}} \\
\beta_{\text{alter}[k]} &= \tilde{\beta}_{\text{alter}[k]} - \frac{1}{K}\sum_{j=1}^{K} \tilde{\beta}_{\text{alter}[j]} \\
\sigma_{\text{teilnehmende}} &\sim \text{Exponential}(1) \\
\beta_{\text{teilnehmende}[p]} &\sim \mathcal{N}\left(\beta_{\text{alter}[k]}, \sigma_{\text{teilnehmende}}\right) \\
\tilde{\mu}_i &= \beta_{\text{teilnehmende}[p[i]]}
\end{aligned}$$

Der Hyperparameter $\beta_{\text{alter}[k]}$ modelliert den durchschnittlichen personenspezifischen Bias pro Altersgruppe und wird unter Verwendung einer unzentrierten Parametrisierung geschätzt sowie anschliessend über die Gruppe K zentriert. Dieser zentrierte Gruppeneffekt dient im nächsten Schritt als Erwartungswert für die Normalverteilung, aus der die individuellen Parameter $\beta_{\text{teilnehmende}[p]}$ gezogen werden. Einzelne Teilnehmende werden somit als Abweichungen von dem durchschnittlichen Bias ihrer jeweiligen Altersgruppe $k[p]$ modelliert.

Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ für Modell M8.1 (*Base+Alter*) entspricht direkt dem personenspezifischen Bias $\beta_{\text{teilnehmende}[p[i]]}$ der Person, welche die Bewertung i abgegeben hat.

Modell M8.2 (*Full+Alter*) integriert diese hierarchische Strukturierung des personenspezifischen Bias in das Modell M4, um die Gruppenunterschiede unter gleichzeitiger Kontrolle der stimulusspezifischen Parameter (*Dialekt, Referenzstimme, Satz*) und der individuellen Annotationssicherheit (a_i) zu verifizieren:

$$\tilde{\mu}_i = \beta_{\text{dialekt}[d[i]]} + \beta_{\text{stimme}[s[i]]} + \beta_{\text{satz}[u[i]]} + \beta_{\text{teilnehmende}[p[i]]} + (\beta_{\text{sicherheit}[p[i]]} \cdot a_i)$$

Die Überführung in den finalen Prädiktor μ_i erfolgt auch in dieser Modellgruppe durch die globale Zentrierung über alle N Gesamtbewertungen.

5.2.8 Modellgruppe M9: Dialektexposition

Die Modellgruppe M9 modelliert den Einfluss der geografischen Dialektexposition der Teilnehmenden auf deren Bewertungsverhalten. Es wird untersucht, ob Personen, die über längere Zeit in bestimmten Kantonen wohnhaft waren, eine veränderte Wahrnehmung oder Sensitivität für die dort gesprochenen Dialekte aufweisen.

Aus den Formularangaben zu den historischen Aufenthaltsorten wird pro teilnehmender Person und Dialekt ein relativer Expositionswert berechnet. Dieser stellt den verhältnismässigen Anteil an der Gesamtexposition dar. Die Kantone werden hierzu über ein fixes Mapping den entsprechenden Dialektregionen zugeordnet (*siehe Tabelle 4*).

Tabelle 4: Kantone zu Dialektmapping. Zuordnung der Schweizer Kantone (inklusive Abkürzung) zu den definierten Dialektklassen (z. B. ch_be, ch_zh). Die Spalte «Gewicht» weist die statische Gewichtung (1) aus. Die Zeile «OTHER» fängt übrige Einträge ohne spezifische Zuweisung («none») auf.

Abk.	Kanton	Dialektzuweisung	Gewicht
BE	Bern	ch_be	1
SO	Solothurn	ch_be	1
BS	Basel-Stadt	ch_bs	1
BL	Basel-Landschaft	ch_bs	1
VS	Wallis	ch_vs	1
ZH	Zürich	ch_zh	1
AG	Aargau	ch_zh	1
LU	Luzern	ch_in	1
OW	Obwalden	ch_in	1
NW	Nidwalden	ch_in	1
UR	Uri	ch_in	1
ZG	Zug	ch_in	1
SZ	Schwyz	ch_in	1
GL	Glarus	ch_in	1
GR	Graubünden	ch_gr	1
SH	Schaffhausen	ch_os	1
TG	Thurgau	ch_os	1
SG	St. Gallen	ch_os	1
AR	Appenzell Ausserrhoden	ch_os	1
AI	Appenzell Innerrhoden	ch_os	1
–	OTHER	none	1

Die Aufenthaltsdauer in Jahren $E_{[p,d]}$ einer Person p in einer Dialektregion d wird aggregiert. Aufenthaltsorte in der Kategorier OTHER (*andere Kantone oder das Ausland*) werden keinem spezifischen Schweizer Dialekt zugewiesen. Da die Werte jedoch anschliessend normiert werden, reduzieren diese Aufenthalte die relative Exposition für die untersuchten Dialektregionen, was den tatsächlichen sprachlichen Hintergrund präzise abbildet. Die mathematische Normierung zur Berechnung der relativen Exposition $\tilde{E}_{[p,d]}$ erfolgt über die Gesamtdauer aller Aufenthaltsjahre T_p einer Person:

$$T_p = \sum_{j}^{D} E_{[p,j]}$$

$$\tilde{E}_{[p,d]} = \frac{E_{[p,d]}}{T_p}$$

Das Modell M9.1 (*Base+Expo*) verwendet diesen normierten Expositionswert analog zu den vorherigen Modellen als globalen Haupteffekt (*Fixed Effect*) mit einem schwach informativen Normalverteilungs-Prior:

$$\beta_{exposition} \sim \mathcal{N}(0,1)$$

Der unzentrierte lineare Prädiktor $\tilde{\mu}_i$ für eine spezifische Bewertung i greift über die Indizes der zugehörigen Person $p[i]$ und des präsentierten Dialekts $d[i]$ interaktiv auf die Expositionsmatrix zu:

$$\tilde{\mu}_i = \beta_{\text{dialekt}[d[i]]} + \beta_{\text{teilnehmende}[p[i]]} + (\beta_{exposition} \cdot \tilde{E}_{[p[i],d[i]]})$$

Das Modell M9.2 (*Full+Expo*) erweitert die Kontrollstruktur aus Modell M4 um diesen Expositionsprädiktor. Dadurch kann der isolierte Effekt der geografischen Prägung unter gleichzeitiger statistischer Kontrolle aller stimulusspezifischen Parameter (*Dialekt, Referenzstimme, Satz*) sowie der individuellen Annotationssicherheit (a_i) zu verifizieren:

$$\tilde{\mu}_i = \beta_{\text{dialekt}[d[i]]} + \beta_{\text{stimme}[s[i]]} + \beta_{\text{satz}[u[i]]} + \beta_{\text{teilnehmende}[p[i]]} + (\beta_{\text{sicherheit}[p[i]]} \cdot a_i) + (\beta_{exposition} \cdot \tilde{E}_{[p[i],d[i]]})$$

Die Überführung in den finalen Prädiktor μ_i erfolgt auch in dieser Modellgruppe durch die globale Zentrierung über alle N Gesamtbewertungen.

6 Resultate

In den folgenden Unterkapiteln werden die Einflüsse verschiedener Prädiktoren auf die Bewertung mithilfe der entwickelten hierarchischen Regressionsmodelle untersucht. Als Basis für die Auswahl und Interpretation der Effekte dient der vorausgehende Modellvergleich.

6.1 Modellverifikation und Prädiktionsgüte

Tabelle 5 fasst alle entwickelten Modelle inklusive ihrer Konvergenzdiagnostiken und LOO-Werte zusammen. Dabei weisen höhere LOO-Werte auf eine bessere prädiktive Modellgüte hin.

Alle Modelle erfüllen die in der Literatur definierten Schwellenwerte für die Konvergenz: Jedes Modell weist einen $\hat{R}$-Wert von unter 1,01 sowie eine effektive Stichprobengrösse (*min. ESS*) von weit über 400 auf. Das Modell M5.1 (*Base+Heim*) wies während des Samplings jedoch sehr hohe Divergenzen auf. Dies macht die Schätzungen, insbesondere für die Analyse des Einflusses des Heimdialekt, unzuverlässig. Aus diesem Grund wird das Modell M5.1 (*Base+Heim*) von den nachfolgenden inhaltlichen Interpretationen ausgeschlossen.

Teilt man die Modelle anhand ähnlicher LOO-Werte in Gruppen mit vergleichbarer prädiktiver Leistung ein, lassen sich folgende systematische Zusammenhänge ableiten:

- **Beitrag der Selbsteinschätzung (*Dialektidentifikation*):** Das Hinzufügen der Selbsteinschätzung (*Vergleich von M4, Full mit M6.2, Full+Selbst*) führt zu einer spürbaren Erhöhung der prädiktiven Modellgüte. Dies deutet auf einen relevanten Beitrag des Effekts hin.
- **Geringer Mehrwert soziodemografischer und externer Faktoren:** Die Modellgüte der Modelle, die zusätzliche Effekte für den Heimdialekt, die Dialektexposition, das Alter oder die Kalibrierungscores beinhalten (*M5.2 bis M9.2*), unterscheidet sich nur marginal. Keiner dieser Parameter führt zu einer substanziellen Verbesserung der Vorhersageleistung gegenüber dem Modell, welches lediglich die Sicherheit berücksichtigt (*M4, Full*).
- **Relevanz der Annotationssicherheit:** Der Güteunterschied zwischen den Modellen mit und ohne den Annotationssicherheitsparameter ist besonders deutlich ausgeprägt (*Vergleich von M3, Base+Stimulus mit M4, Full*). Dies belegt den starken Beitrag der subjektiven Sicherheit zur Vorhersageleistung des Modells.
- **Bedeutung der stimulusspezifischen Effekte:** Modelle, die die hierarchische Struktur der Stimuli nur unvollständig abbilden (*z. B. nur den Dialekt oder nur die Referenzstimme*), zeigen eine deutlich schlechtere Vorhersagegüte. Dies spricht dafür, dass alle drei Komponenten (*Dialekt, Referenzstimme und Satz*) wesentlich zur Vorhersagekraft der Modelle beitragen.

Für die detaillierte Analyse der spezifischen Effekte in den folgenden Abschnitten wird das Modell mit der höchsten prädiktiven Gesamtgüte verwendet: M6.2.

Tabelle 5: Bayesscher Modellvergleich und Konvergenzdiagnostik der Regressionsmodelle. Modelle absteigend sortiert nach prädiktiver Güte gemäss Leave-One-Out Cross-Validation (LOO). Spalten umfassen den LOO-Wert, die Differenz zum besten Modell (Diff-LOO), den Standardfehler der Differenz (DSE) sowie die MCMC-Konvergenzprüfmasse RHAT und die minimale effektive Stichprobengrösse (Min. ESS Bulk/Tail).

Modell	LOO	Diff-LOO	DSE	RHAT	Min. ESS (Bulk/Tail)
M6.2 (Full+Selbst)	-11977.46	0	0	1.0054	956
M4 (Full)	-11991.33	13.86	5.46	1.0042	1262
M5.2 (Full+Heim)	-11991.73	14.27	5.25	1.0051	1213
M7.2 (Full+Kalib)	-11992.01	14.55	5.73	1.0056	1283
M8.2 (Full+Alter)	-11992.19	14.73	5.48	1.0045	1201
M9.2 (Full+Expo)	-11992.72	15.26	5.34	1.0044	1095
M3 (Base+Stimulus)	-12454.06	476.60	30.01	1.0080	1148
M2.3 (Base+Satz)	-12674.20	696.74	35.30	1.0052	1619
M7.1 (Base+Kalib)	-12860.32	882.86	38.98	1.0060	1115
M9.1 (Base+Expo)	-12863.32	886.00	38.85	1.0054	1379
M6.1 (Base+Selbst)	-12864.43	886.97	38.69	1.0061	1246
M2.1 (Base+Dialekt)	-12866.34	888.87	38.78	1.0061	1127
M2.2 (Base+Stimme)	-12948.78	971.31	40.43	1.0055	1155
M5.1 (Base+Heim)	-13006.28	1028.82	41.29	1.0056	916
M8.1 (Base+Alter)	-13009.66	1032.20	41.38	1.0074	1007
M1 (Base)	-13009.76	1032.30	41.38	1.0038	1727

6.2 Relative Relevanz der Modellkomponenten

Zur Quantifizierung und vergleichenden Einordnung der Effektstärken einzelner Modellkomponenten werden diese über ihre Varianzanteile am gesamten linearen Prädiktor mittels Varianzanteilskoeffizienten (*Variance Partition Coefficient, VPC*) beschrieben. Da die Modellkomponenten nicht immer unabhängig sind und teilweise in Form von Interaktionstermen eingehen, ist die resultierende Varianz nicht strikt additiv zerlegbar. Die VPCs sind daher als approximative marginale Beiträge und nicht als exakte Varianzdekomposition zu interpretieren.

In Abbildung 24 sind die VPCs der einzelnen Effekte dargestellt. Daraus lassen sich folgende Kernaussagen ableiten:

- **Dominanz des personenspezifischen Bias und der Sicherheit:** Die beiden Effekte des personenspezifischen Bias und der Annotationssicherheit weisen in allen Modellen, in denen sie berücksichtigt werden, deutlich höhere Varianzanteile auf als die übrigen Prädiktoren. Dies deutet darauf hin, dass die Variabilität in der Dialektbewertung primär durch personenspezifische Unterschiede und das subjektive Vertrauen in die eigene Bewertung erklärt wird.
- **Interdependenz der Prädiktoren:** Die Effekte des personenspezifischen Bias und der Annotationssicherheit weisen eine gewisse Kollinearität auf. Beim Hinzufügen der Annotationssicherheit im Modell M4 (*Full*) im Vergleich zu M3 (*Base+Stimulus*) reduziert sich die Gesamtsumme der übrigen Varianzanteile, da die Sicherheit einen Teil der zuvor unerklärten oder personenspezifischen Varianz absorbiert.
- **Geringe Relevanz sekundärer Faktoren:** Die Effekte Heimdialekt, Selbsteinschätzung (*Dialektidentifikation*), Kalibrierungsscore und Dialektexposition besitzen in allen Modellarchitekturen nur sehr geringe Varianzanteile. Die durch sie transportierte Information wird weitestgehend durch die übergeordnete personenspezifische Varianz (*Random Intercepts*) abgedeckt.

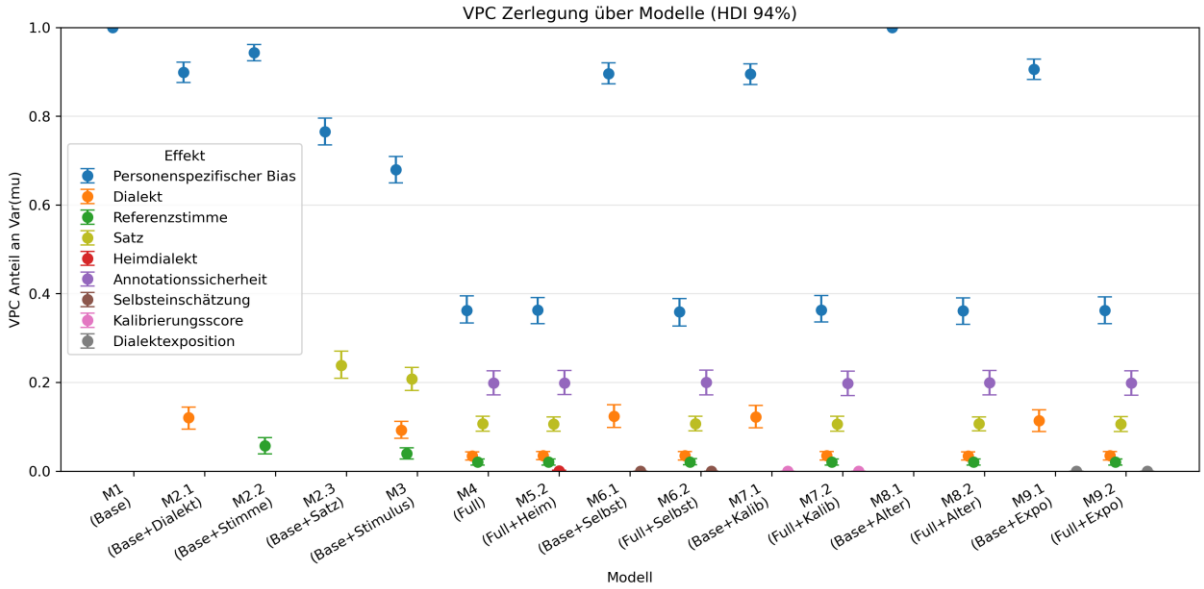

Abbildung 24: Varianzanteilskoeffizienten (VPC). Relativer Anteil der Zufallseffekte und Prädiktoren an der Gesamtvarianz von μ (Y-Achse) für die einzelnen Modelle (X-Achse). Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI). Farben codieren die modellierten Effekte.

Abbildung 25 veranschaulicht das relative Verhältnis der stimulusspezifischen Effekte zueinander, wenn die dominanten personenspezifischen Parameter ausgeblendet werden (*Bias und Annotationssicherheit*). Es zeigt sich der folgende Modellübergreifende Trend: Der VPC der einzelnen Sätze ist signifikant höher als der aller anderen verbleibenden Komponenten. Die sprachliche Struktur oder der Inhalt der fünf Sätze übt somit einen deutlich stärkeren Einfluss auf die Bewertung aus als die Referenzstimme oder der Dialekt.

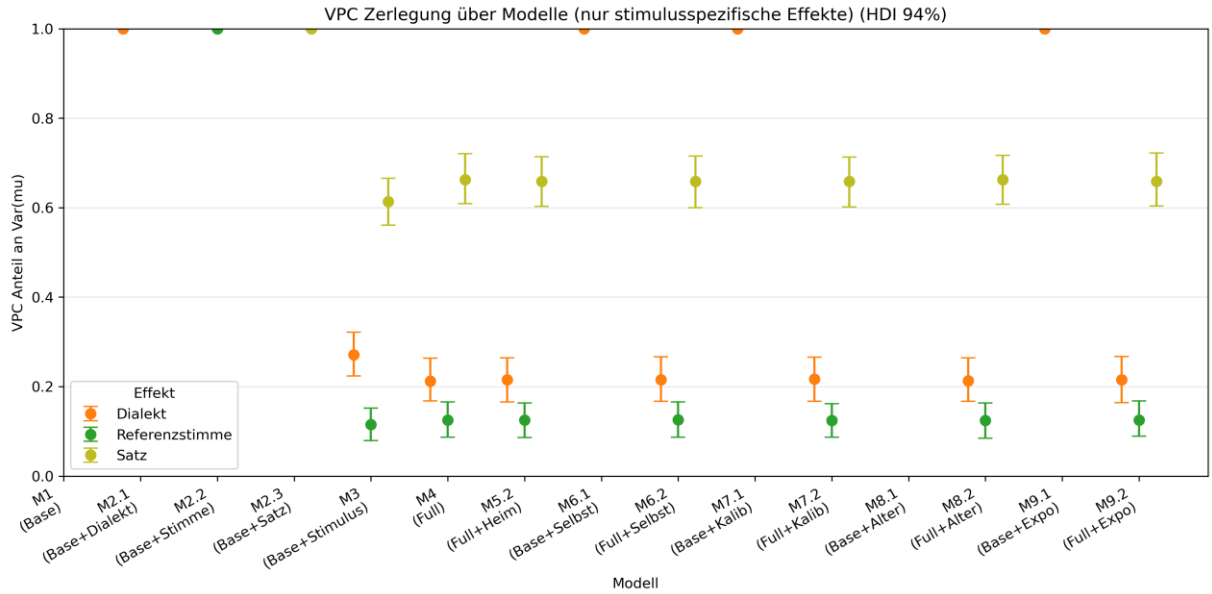

Abbildung 25: Varianzanteilskoeffizienten (VPC (nur mit stimulusspezifischen Effeken). Relativer Anteil der Prädiktoren an der bereinigten Varianz von μ (Y-Achse) über die Modellstufen (X-Achse). Die Varianz wurde zuvor um alle anderen Effekte bereinigt. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI). Farben codieren die modellierten Effekte.

6.3 Analyse der stimulusspezifischen Effekte

Nachdem die relative Relevanz der einzelnen Modellkomponenten über die Varianzanteile bestimmt wurde, folgt eine detaillierte Analyse der stimulusspezifischen Einflussfaktoren (*Dialekt, Referenzstimme, Satz*). Sämtliche Parameterschätzungen und Prädiktionen beziehen sich, sofern nicht anders ausgewiesen, auf das prädiktiv stärkste Modell M6.2 (*Full+Selbst*).

6.3.1 Dialekteffekte

Die Untersuchung der Dialekteffekte über alle Modelle hinweg offenbart eine Robustheit gegenüber veränderten Modellkonfigurationen (*siehe Abbildung 26*). Die geschätzten Regressionskoeffizienten (*Beta-Werte*) zeigen jedoch eine systematische Verschiebung der posterioren Verteilungen, sobald die Annotationssicherheit als Kontrollvariable einbezogen wird. Dieser Effekt wird besonders beim Dialekt Innerschweiz (*IN*) sichtbar: Dessen negativer Einfluss auf die Bewertung fällt nach der Adjustierung für die Sicherheit signifikant schwächer aus. Dies deckt sich mit der Beobachtung, dass Stimuli aus der Innerschweiz im Durchschnitt mit einer geringeren Sicherheit evaluiert wurden und ein Teil dieser Varianz somit über den Sicherheitseffekt absorbiert wird.

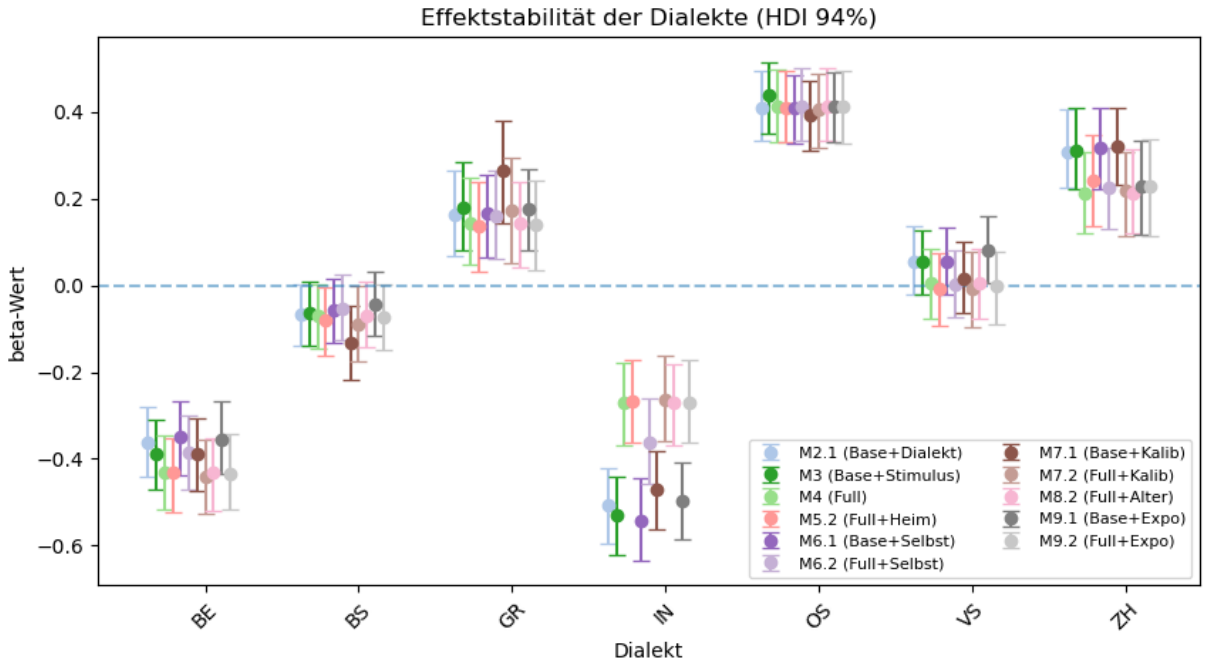

Abbildung 26: Effektstabilität der Dialekte. Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für die sieben Dialektregionen (X-Achse) über elf Modelle. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

Hinsichtlich der Bewertungsverteilungen im Modell M6.2 (*Full+Selbst*) lassen sich die untersuchten Dialekte in zwei kontrastierende Gruppen unterteilen (*siehe Abbildung 27*):

- **Negativer Bewertungstrend:** Für die Dialekte Bern (*BE*) und Innerschweiz (*IN*) werden im Vergleich zu den anderen Dialekten systematisch geringere Wahrscheinlichkeiten für die Vergabe hoher Bewertungskategorien prognostiziert.
- **Positiver Bewertungstrend:** Die Dialekte Ostschweiz (*OS*), Zürich (*ZH*) und Graubünden (*GR*) erzielen eine signifikant höhere Wahrscheinlichkeit, in den oberen Skalensegmenten bewertet zu werden.

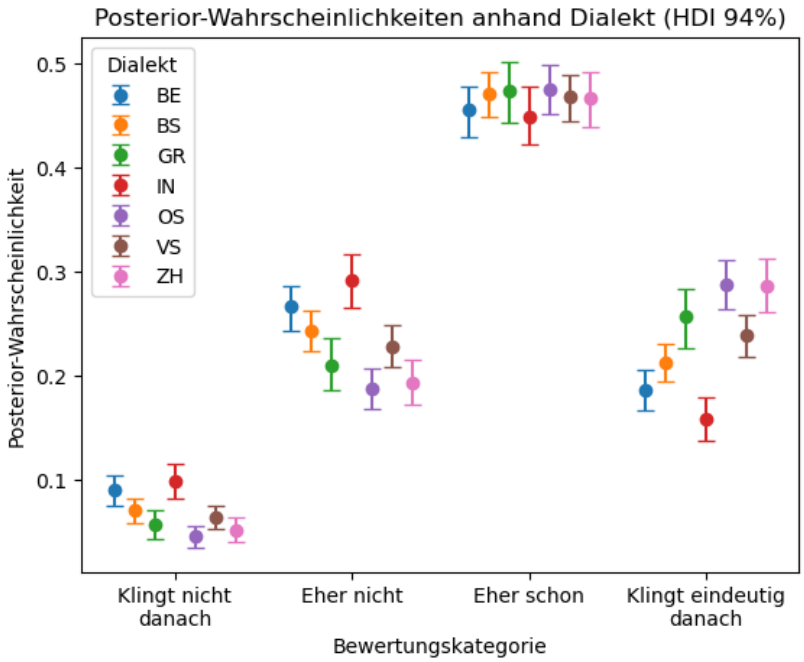

Abbildung 27: Posterior-Wahrscheinlichkeiten anhand Dialekt (M6.2). Geschätzte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Bewertungskategorien 0 bis 3 (X-Achse). Punkte markieren posteriore Mediane für die sieben Dialektregionen (farblich codiert), vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI).

Die Betrachtung der marginalisierten Posterior-Wahrscheinlichkeiten, bei denen sämtliche verbleibenden Modellparameter auf den Erwartungswert Null gesetzt werden, isoliert den rein durch den Dialekt erklärten Varianzanteil für eine durchschnittliche Person (*siehe Abbildung 28*). Die Marginalisierung verdeutlicht erneut, dass bei Bern und der Innerschweiz ein beträchtlicher Teil der Bewertungsvarianz im vollständigen Modell durch personenspezifische Effekte, wie die Annotationssicherheit, getragen wird.

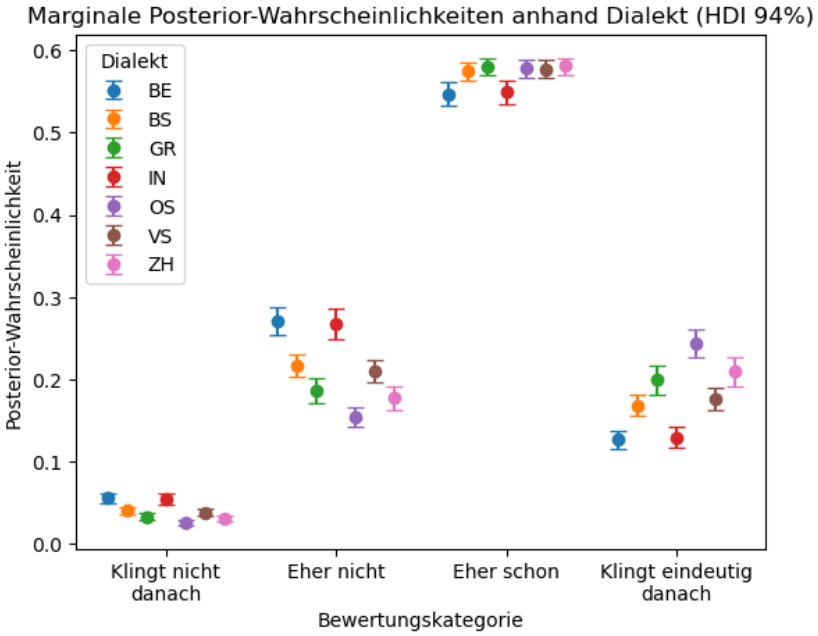

Abbildung 28: Marginale Posterior-Wahrscheinlichkeiten anhand Dialekt (M6.2). Marginalisierte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Antwortstufen 0 bis 3 (X-Achse). Punkte markieren posteriore Mediane für die sieben Dialektregionen unter Marginalisierung aller anderen Modellparameter. Vertikale Fehlerbalken bilden das 94%-Highest-Density-Interval (HDI) ab.

6.3.2 Referenzstimmeffekte

Im Gegensatz zu den Dialekten erweist sich der Einfluss der einzelnen Referenzstimmen über alle Modellstrukturen hinweg als mathematisch vollkommen stabil (*siehe Abbildung 29*). Weder das Hinzufügen der Annotationssicherheit noch andere personenspezifischer Variablen führen zu einer merklichen Verschiebung der posterioren Verteilungen.

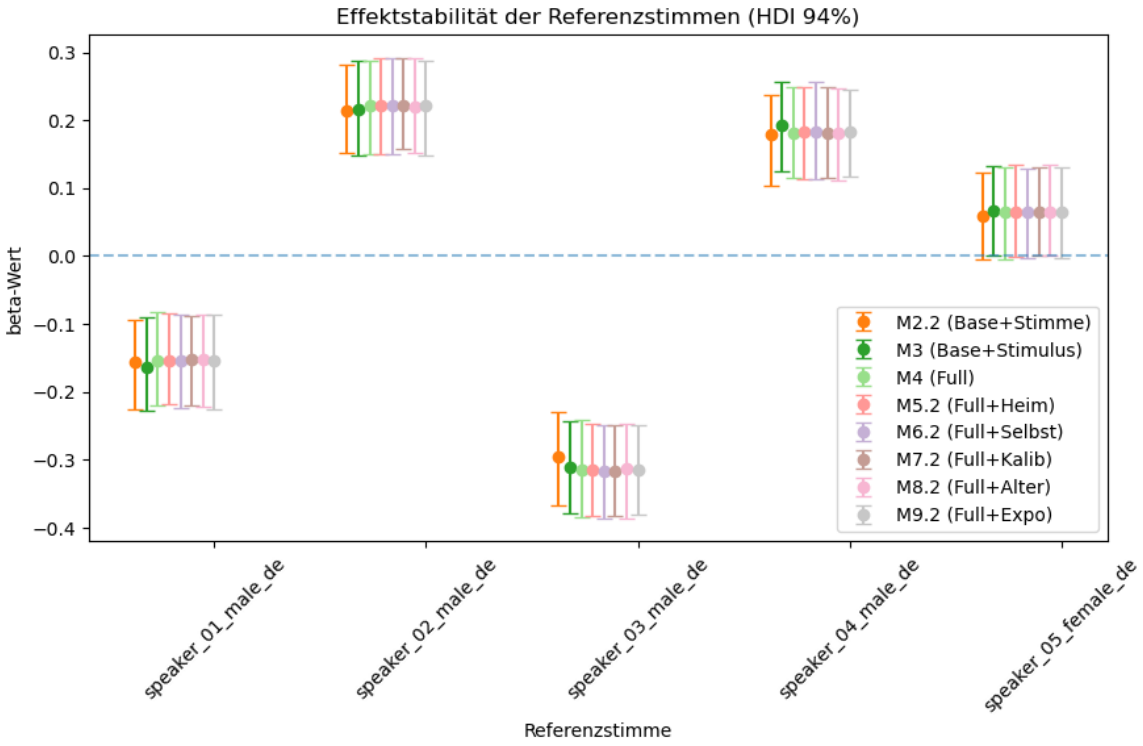

Abbildung 29: Effektstabilität der Referenzstimmen. Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für die fünf Referenzstimmen (X-Achse) über acht Modelle. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

Die Analyse des Modells M6.2 (*Full+Selbst*) zeigt eine klare Hierarchie innerhalb der Referenzstimmen (*siehe Abbildung 30*): Die Stimmen «speaker_01_male_de» und «speaker_03_male_de» werden über das gesamte Sample hinweg systematisch schlechter bewertet als die restlichen drei Stimmen.

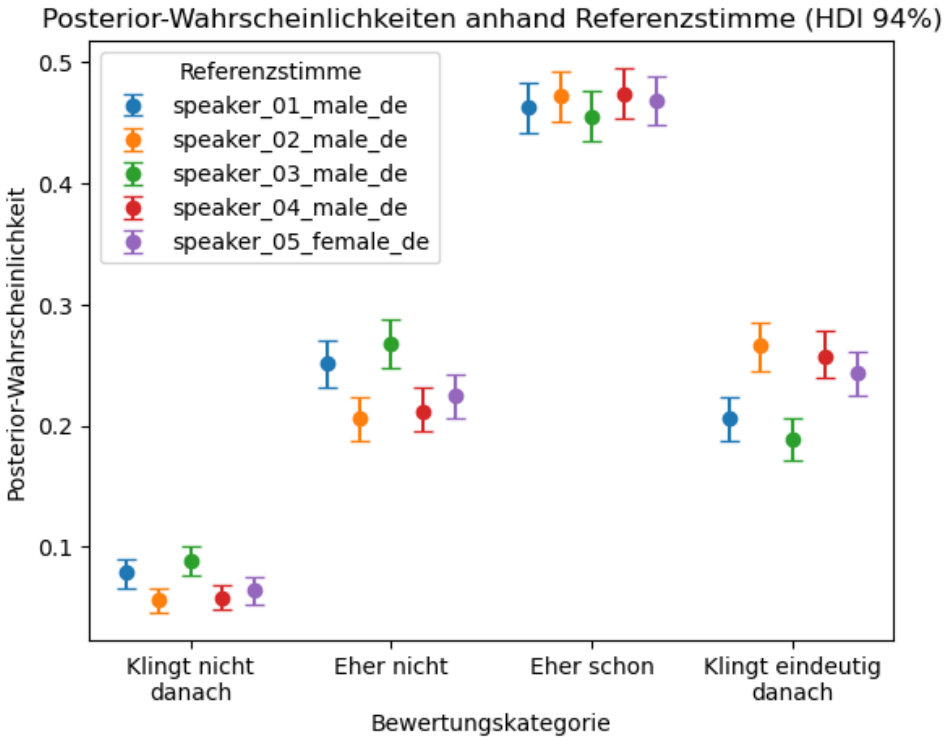

Abbildung 30: Posterior-Wahrscheinlichkeiten anhand Referenzstimme (M6.2). Geschätzte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Bewertungskategorien 0 bis 3 (X-Achse). Punkte markieren posteriore Mediane für die fünf Referenzstimmen (farblich codiert), vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI).

Die marginalisierten Posterior-Wahrscheinlichkeiten (*siehe Abbildung 31*) weisen im Vergleich zur vollständigen Prädiktion – abgesehen von der Reduktion der Unsicherheitsintervalle durch Fixierung der Zufallseffekte – keine informationellen Unterschiede auf.

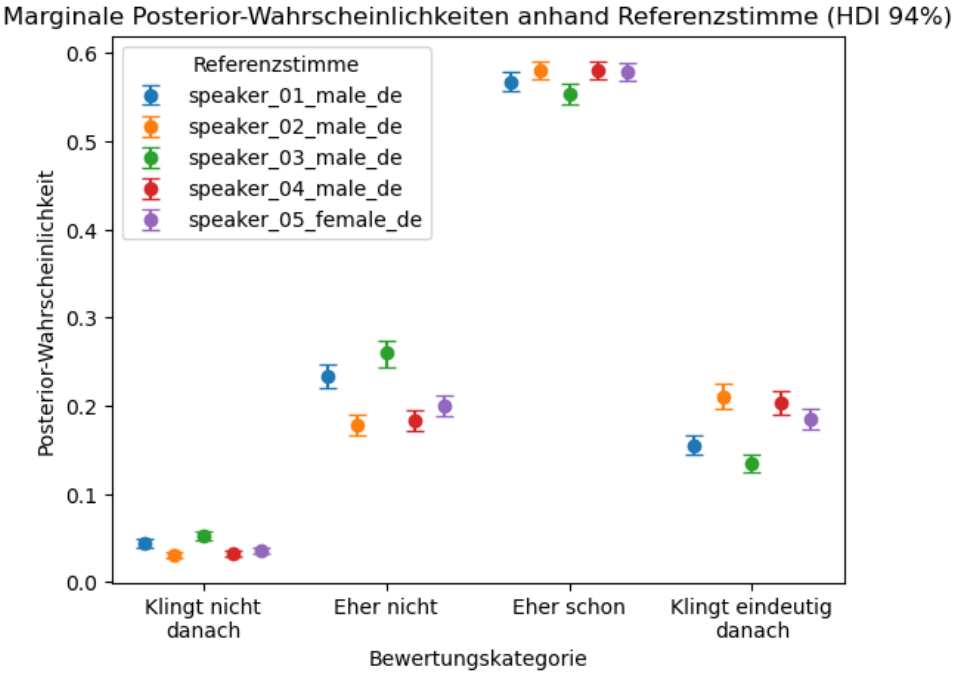

Abbildung 31: Marginale Posterior-Wahrscheinlichkeiten anhand Referenzstimme (M6.2). Marginalisierte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Antwortstufen 0 bis 3 (X-Achse). Punkte markieren posteriore Mediane für die fünf Referenzstimmen unter Marginalisierung aller anderen Modellparameter. Vertikale Fehlerbalken bilden das 94%-Highest-Density-Interval (HDI) ab.

Um potenzielle Wechselwirkungen zwischen Referenzstimme und Dialekt zu prüfen, wurden kontrafaktische Posterior-Wahrscheinlichkeiten berechnet (*siehe Abbildung 32*). Die Ergebnisse belegen eine strikte Unabhängigkeit der beiden Faktoren: Innerhalb jeder Referenzstimme manifestieren sich dieselben Dialekteffekte in identischer relativer Ausprägung. Dies deckt sich mit der additiven Varianzzerlegung der Basismodelle (*bis einschliesslich M3, Base+Stimulus*), bei denen die Summe der marginalen VPCs nahezu eins ergibt. Das schliesst verborgene Interaktionseffekte empirisch aus. Die Stimmencharakteristik beeinflusst das Rating somit rein als additiver Shift.

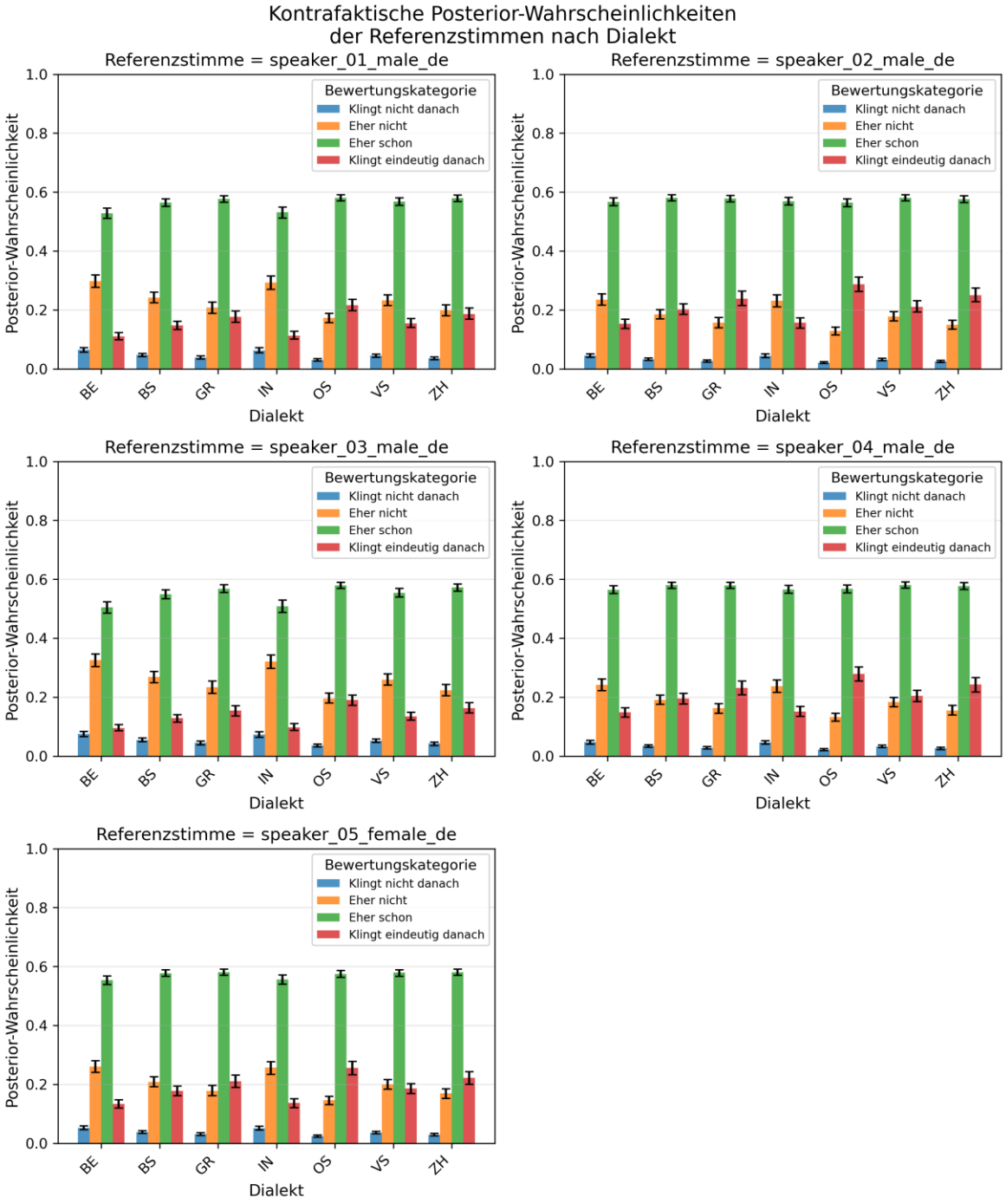

Abbildung 32: Kontrafaktische Posterior-Wahrscheinlichkeiten der Referenzstimmen nach Dialekt (M6.2). Geschätzte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Antwortstufen 0 bis 3 (farblich codiert) in Abhängigkeit der sieben Dialektregionen (X-Achse). Facettenplots separieren nach den fünf Referenzstimmen zur Abbildung sprecherspezifischer Effekte. Fehlerbalken markieren die Schätzungsunsicherheit.

6.3.3 Satzeffekte

Der stärkste stimulusspezifische Prädiktor im Modell resultiert aus den unterschiedlichen Sätzen. Entsprechend ihrer dominanten Rolle in der VPC-Analyse zeigt sich die Parameterschätzung für die fünf Sätze über alle Modelle hinweg als äusserst stabil und weist sehr enge Unsicherheitsintervallen auf (*siehe Abbildung 33*).

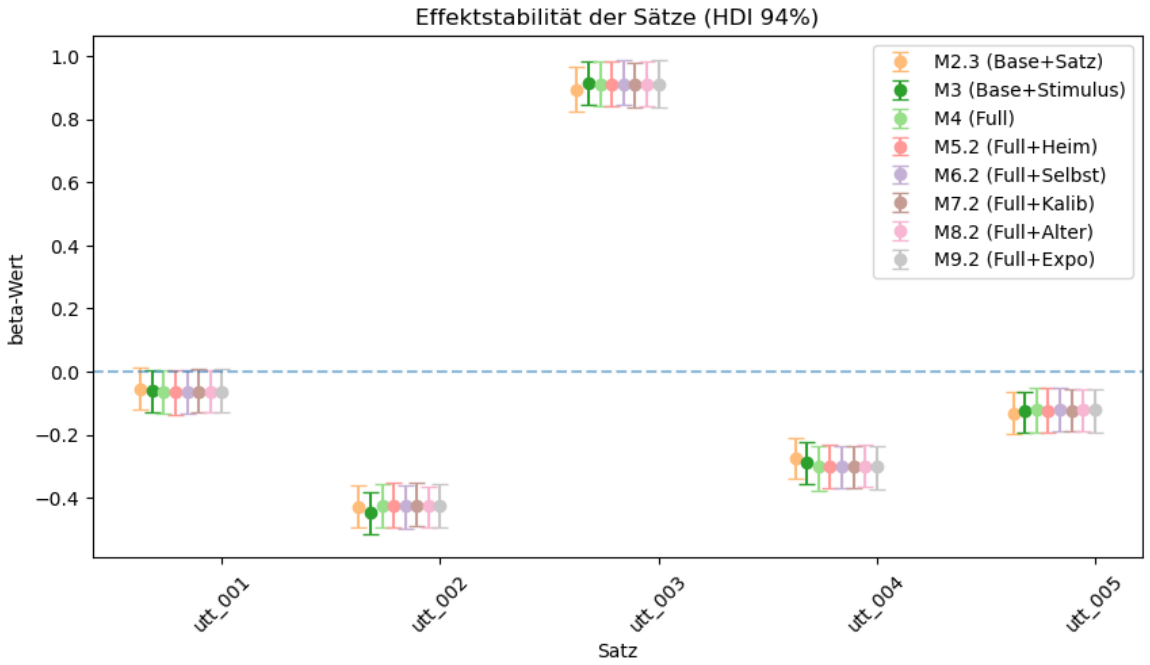

Abbildung 33: Effektstabilität der Sätze. Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für die fünf Sätze (X-Achse) über acht Modelle. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

Die marginalisierten Verteilungen des Modells M6.2 (*Base+Selbst*) offenbaren signifikante Diskrepanzen im resultierenden Bewertungsverhalten (*siehe Abbildung 34*):

- **utt_003:** Dieser Stimulus verzeichnet im Vergleich zum Gesamtsample eine deutlich erhöhte Wahrscheinlichkeit für Spitzenbewertungen. Dies könnte darauf zurückzuführen sein, dass der Textinhalt mehr dialektrelevante Schlüsselwörter enthält als die anderen Sätze.
- **utt_002:** Dieser Textinhalt führt zu einer systematisch schlechteren Einstufung durch die annotierenden Personen.

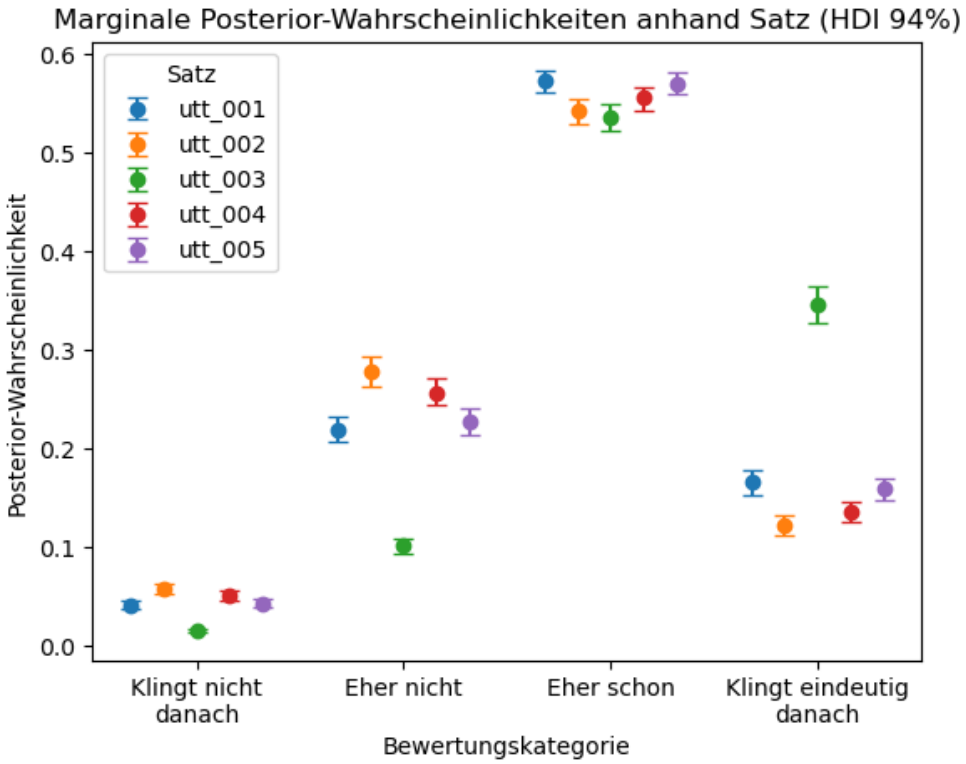

Abbildung 34: Marginale Posterior-Wahrscheinlichkeiten anhand Satz (M6.2). Marginalisierte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Antwortstufen 0 bis 3 (X-Achse). Punkte markieren posteriore Mediane für die fünf Sätze unter Marginalisierung aller anderen Modellparameter. Vertikale Fehlerbalken bilden das 94%-Highest-Density-Interval (HDI) ab.

Analog zur Analyse der Referenzstimmen belegt die kontrafaktische Betrachtung über die Dialekte hinweg (*siehe Abbildung 35*) auch für die Sätze das Fehlen von Interaktionseffekten. Die relativen Bewertungsunterschiede zwischen den Dialekten bleiben von Satz zu Satz konstant. Die syntaktische oder semantische Beschaffenheit des Satzes steuert das Bewertungsverhalten somit massiv und unabhängig vom dargebotenen Dialekt.

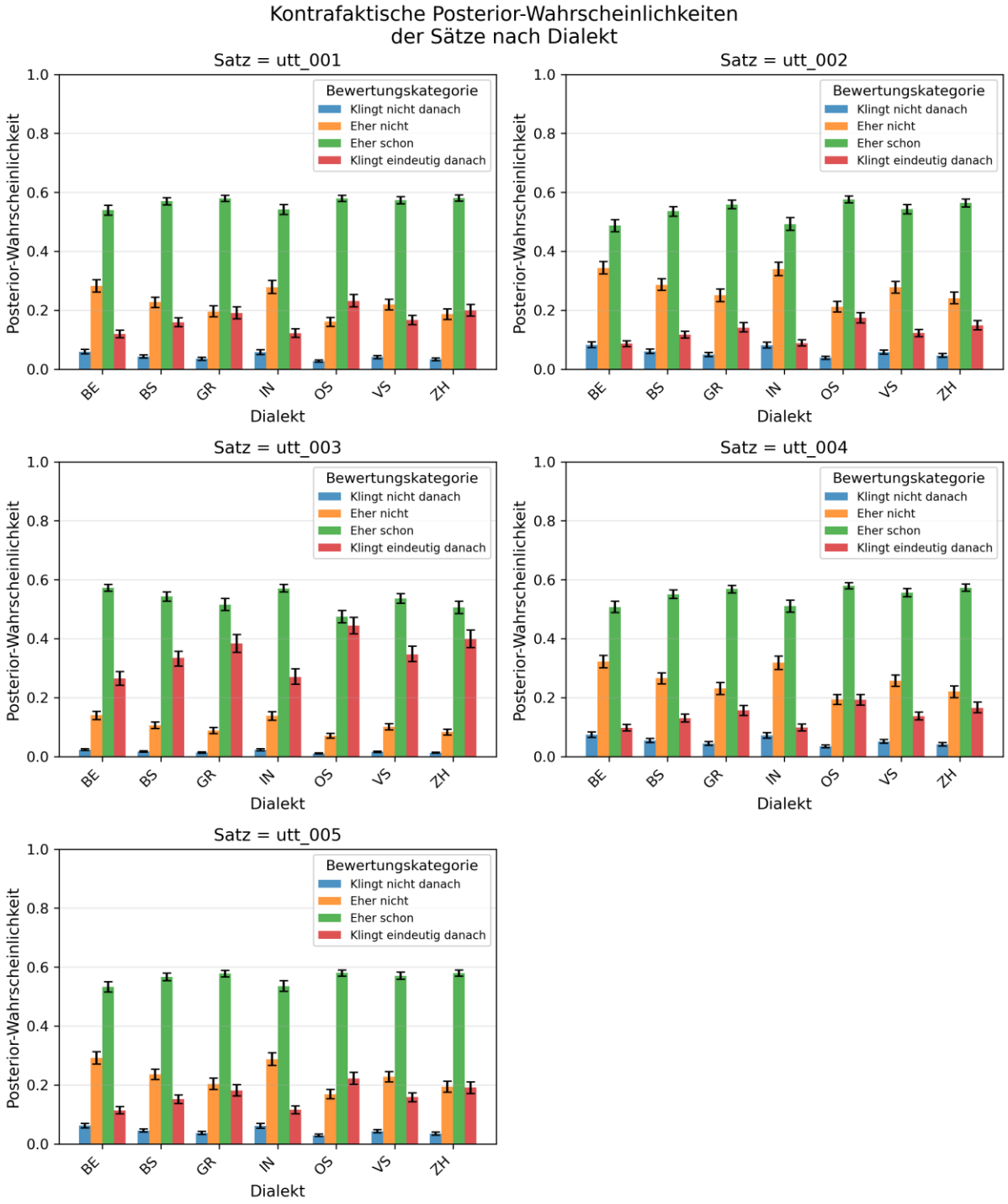

Abbildung 35: Kontrafaktische Posterior-Wahrscheinlichkeiten der Sätze nach Dialekt (M6.2). Geschätzte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Antwortstufen 0 bis 3 (farblich codiert) in Abhängigkeit der sieben Dialektregionen (X-Achse). Facettenplots separieren nach den fünf Sätzen zur Abbildung satzspezifischer Effekte. Fehlerbalken markieren die Schätzungsunsicherheit.

6.4 Analyse der personenspezifischen Effekte

Nach der Evaluation der stimulispezifischen Parameter rückt nun der Einfluss der individuellen Charakteristika der Teilnehmenden in den Fokus. Da die Varianzanteilsanalyse gezeigt hat, dass personenspezifische Unterschiede die Variabilität in den Bewertungen dominieren, werden die entsprechenden Prädiktoren im Folgenden detailliert betrachtet. Die Untersuchung wird dabei systematisch in relevante und vernachlässigbare Effekte unterteilt.

6.4.1 Relevante Effekte: Annotationssicherheit und Selbsteinschätzung

Der markanteste personenspezifische Einflussfaktor zeigt sich in der subjektiven Annotationssicherheit. Wie die Schätzungen über alle Modelle belegen, weist dieser Parameter eine hohe Stabilität auf (*siehe Abbildung 36*). Das Einbetten dieser Variable führt zudem zu einer signifikanten Steigerung der Vorhersagegüte des Gesamtmodells, wie bereits in Kapitel 6.1 gezeigt.

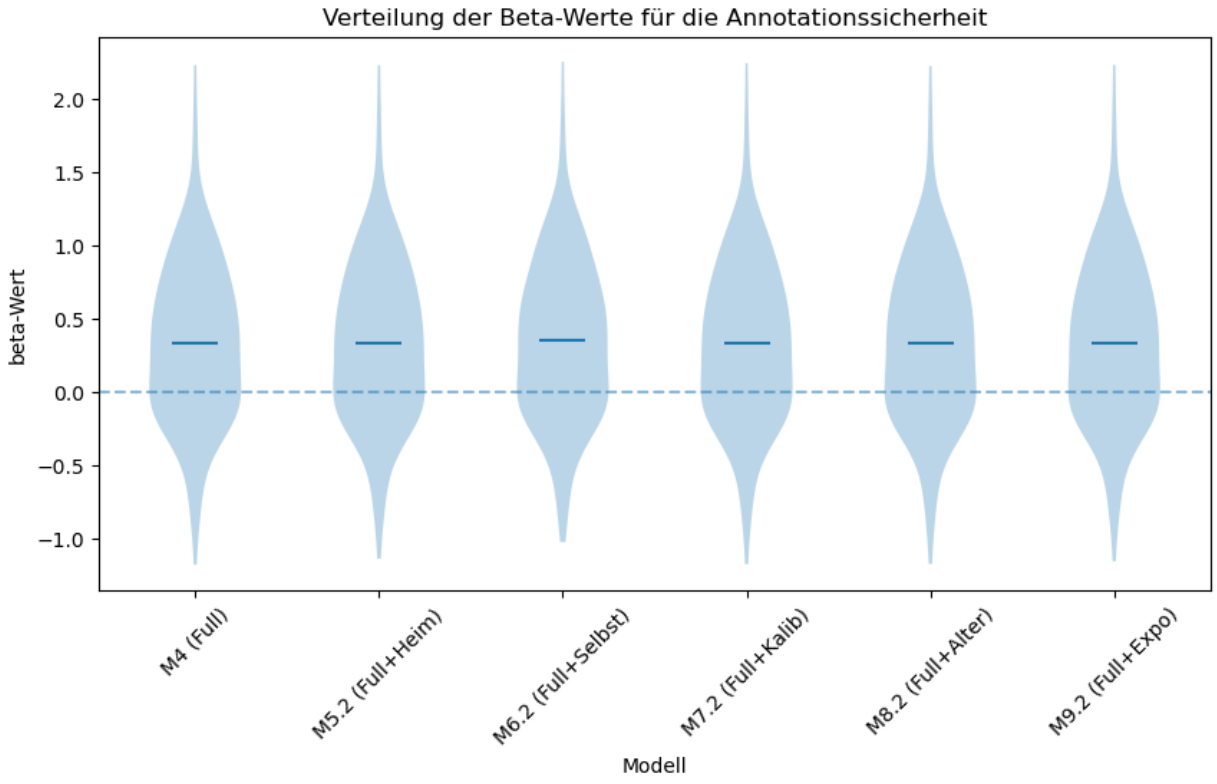

Abbildung 36: Verteilung der Beta-Werte für die Annotationssicherheit. Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für die Annotationssicherheit über sechs Modelle (X-Achse). Im Violin-Plot markieren horizontale Linien die posterioren Mediane, die vertikale Ausdehnung die Dichte und die gestrichelte Linie den Nullpunkt (kein Effekt).

Die Analyse der marginalisierten Posterior-Wahrscheinlichkeiten im Modell M6.2 (*Full+Selbst*) verdeutlicht einen starken linearen Zusammenhang zwischen der erfassten Sicherheit und der Bewertung (*siehe Abbildung 37*):

- **Geringe Annotationssicherheit:** Wenn Teilnehmende eine niedrige Sicherheit angeben, verschiebt sich die Wahrscheinlichkeitsmasse der Schätzungen deutlich in Richtung der unteren Skalensegmente.
- **Hohe Annotationssicherheit:** Mit steigendem Vertrauen in das eigene Urteil erhöht sich die Wahrscheinlichkeit für die Vergabe von Spitzenbewertungen substanziell.

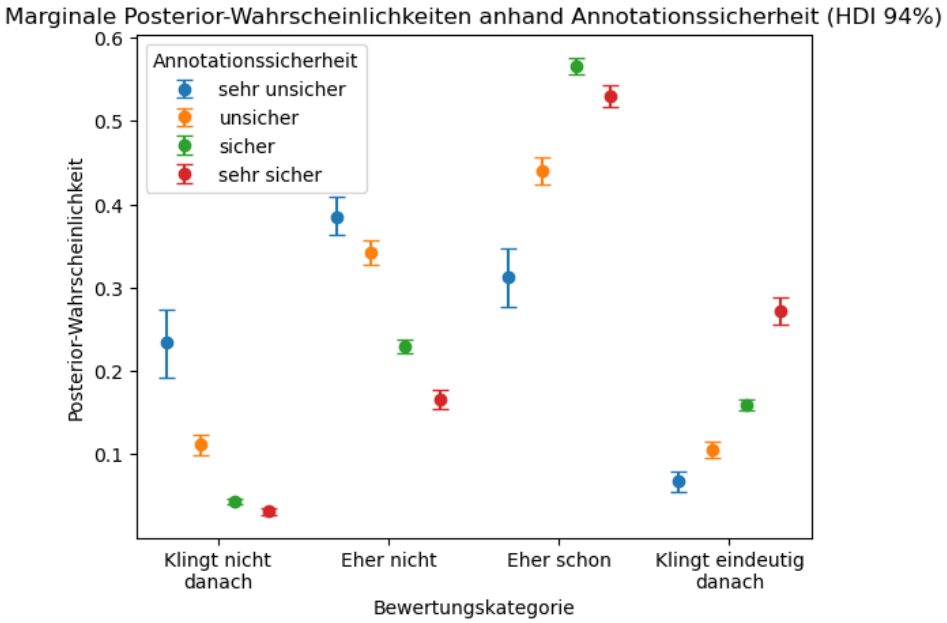

Abbildung 37: Marginale Posterior-Wahrscheinlichkeiten anhand Annotationssicherheit (M6.2). Marginalisierte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Antwortstufen 0 bis 3 (X-Achse). Punkte markieren posteriore Mediane für die vier Sicherheitsstufen unter Marginalisierung aller anderen Modellparameter. Vertikale Fehlerbalken bilden das 94%-Highest-Density-Interval (HDI) ab.

Neben der Annotationssicherheit erweist sich auch die Selbsteinschätzung als valider Prädiktor für das Bewertungsverhalten. Die Schätzungen dieses Parameters zeigen über alle Modellarchitekturen hinweg eine hohe Robustheit (*siehe Abbildung 38*). Der Effekt wird ausgeprägter, wenn für weitere personenspezifische Kovariaten kontrolliert wird. Diese Relevanz deckt sich mit der informationstheoretischen Modellselektion: Der signifikante Zuwachs an prädiktiver Modellgüte von Modell M4 (*Full*) zum Modell M6.2 (*Full+Selbst*) ist primär auf die Integration dieses Effekts zurückzuführen.

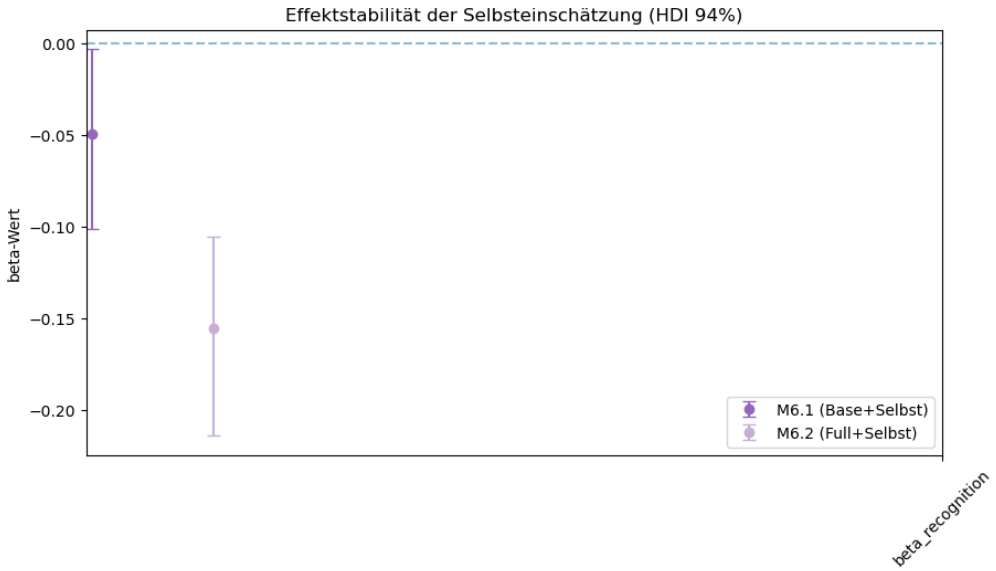

Abbildung 38: Effektstabilität Selbsteinschätzung (Dialektidentifikation) (M6.1, M6.2). Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für den Selbsteinschätzungseffekt über zwei Modelle. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

Die detaillierte Betrachtung der marginalisierten Posterior-Wahrscheinlichkeiten in M6.2 (*Full+Selbst*) unter Kontrolle des personenspezifischen Bias, der Annotationssicherheit sowie der Dialekte (*siehe Abbildung 39*) legt eine gegenläufige Verhaltensdynamik offen: Eine höhere Selbsteinschätzung führt systematisch zu tieferen Bewertungen. Dies liegt nahe, dass Teilnehmende, die den dargebotenen Dialekt vermeintlich oder tatsächlich besser identifizieren, bei der anschliessenden Bewertung einen strengeren Massstab anlegen.

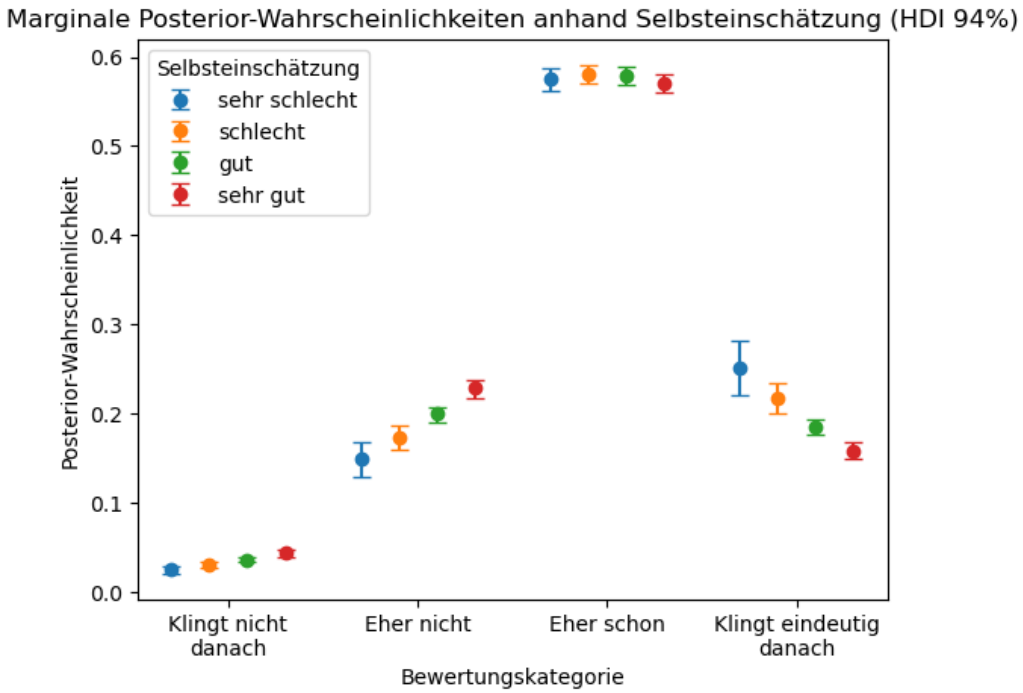

Abbildung 39: Marginale Posterior-Wahrscheinlichkeiten anhand Selbsteinschätzung (Dialektidentifikation) (M6.2). Marginalisierte Posterior-Wahrscheinlichkeiten (Y-Achse) für die vier Antwortstufen 0 bis 3 (X-Achse). Punkte markieren posteriore Mediane für die vier Selbsteinschätzungs-Stufen unter Marginalisierung aller anderen Modellparameter. Vertikale Fehlerbalken bilden das 94%-Highest-Density-Interval (HDI) ab.

6.4.2 Vernachlässigbare Effekte: Heimdialekt, Kalibrierungsscore, Alter, Dialektexposition

Ein zentrales Ergebnis der Modellselektion (*mittels LOO-CV*) ist, dass mehrere soziodemografische, geografische und experimentelle Kontrollvariablen keinen systematischen Erklärungswert für das Bewertungsverhalten besitzen. Diese Faktoren weisen breite Unsicherheitsintervalle auf, die stark um den Nullpunkt streuen, und führen zu keiner signifikanten Verbesserung der prädiktiven Leistung. Aus Gründen der statistischen Sparsamkeit (*Parsimonie*) und zur Vermeidung von Overfitting werden diese Effekte im finalen Modell nicht primär interpretiert, sondern nachfolgend als methodische Null-Effekte dokumentiert.

6.4.2.1 Heimdialekteffekte

Die Überprüfung, ob eine Übereinstimmung zwischen dem eigenen Dialekt der annotierenden Person und dem Dialekt des dargebotenen Stimulus die Bewertung systematisch beeinflusst, liefert keinen Beleg für einen solchen Effekt. Wie die Schätzung im Modell M5.2 (*Full+Heim*) verdeutlicht (*siehe Abbildung 40*), ist der Parameter mit massiver Unsicherheit behaftet. Der informationstheoretische Modellvergleich bestätigt, dass das Hinzufügen dieses Effekts die Vorhersagegüte nicht verbessert. Der Heimdialekt ist übt folglich keinen nennenswerten Einfluss auf das Bewertungsmuster aus.

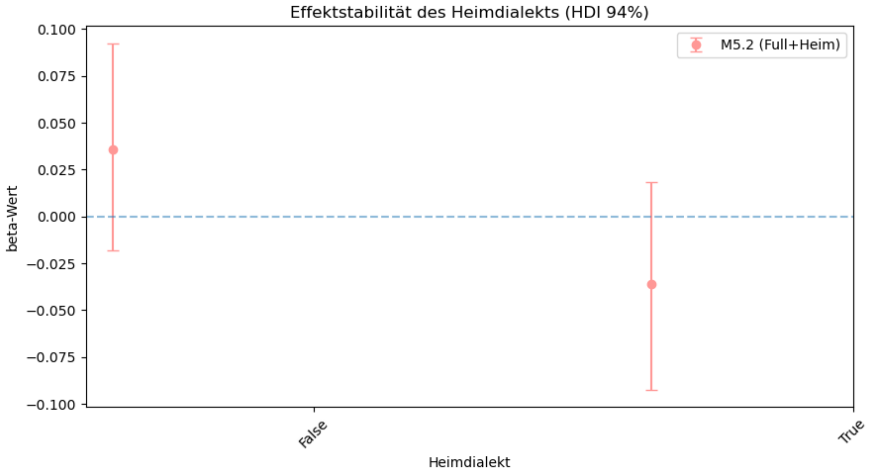

Abbildung 40: Effektstabilität des Heimdialekts (M5, M6). Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für den Heimdialekt-Effekt über ein Modell. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

6.4.2.2 Kalibrierungsscore-Effekte

Die Scores aus der Kalibrierungsphase wurden verwendet, um den globalen Erkennungsgrad der jeweiligen Dialekte statistisch zu kontrollieren. Die Stabilitätsprüfung des entsprechenden Beta-Parameters (*siehe Abbildung 41*) offenbart jedoch, dass dieser aggregierte Effekt verschwindet, sobald personenspezifische Parameter in das Modell integriert werden. Ein prädiktiver Mehrwert des erweiterten Modells M7.2 (*Full+Kalib*) gegenüber dem Basismodell M4 (*Full*) lässt sich nicht nachweisen.

Daraus lässt sich ein zentraler kognitiver Aspekt des Bewertungsprozesses ableiten: Der gruppenbasierte Erkennungsgrad eines Dialekts besitzt keinen eigenständigen Erklärungswert mehr, sobald für die Annotationssicherheit der einzelnen Person kontrolliert wird. Das individuelle Vertrauen in das eigene Urteil im Moment der Bewertung absorbiert die Varianz des vorgelagerten Kalibrierungsscores vollständig und bildet die entscheidende psycholinguistische Kontrollgrösse.

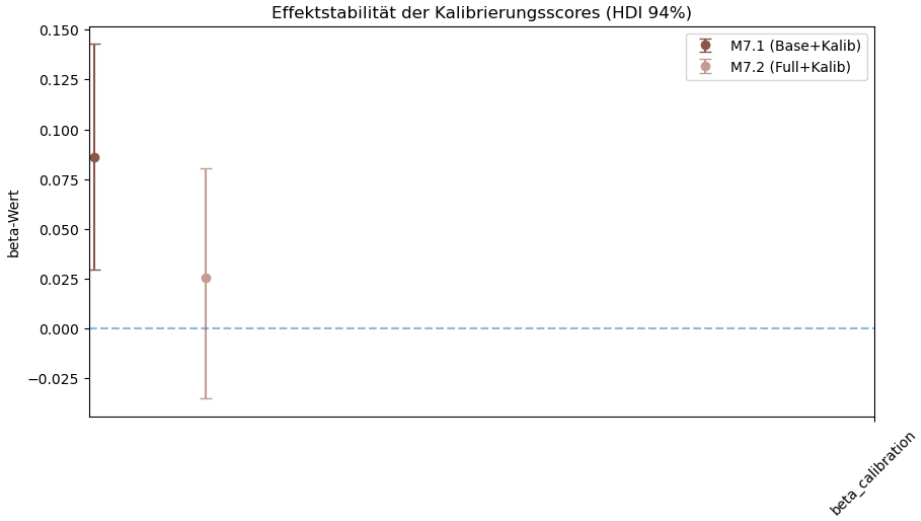

Abbildung 41: Effektstabilität des Kalibrierungsscores (M7.1, M7.2). Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für den Kalibrierungsscore-Effekt über zwei Modelle. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

6.4.2.3 Alterseffekte

Die visuelle Inspektion der Parameterstabilität hinsichtlich des Alters (*siehe Abbildung 42*) zeigt, dass dieser Effekt hochgradig unsicher ist. Innerhalb der Altersgruppen dominieren stark divergierende, individuelle Bewertungsstrategien, weshalb sich kein systematischer Alterseffekt isolieren lässt. Folglich generiert das Modell M8.2 (*Full+Alter*) im Vergleich zum Basismodell M4 (*Full*) nahezu identische LOO-CV-Werte.

An dieser Stelle ist jedoch eine methodische Einschränkung zu berücksichtigen: Die starke Streuung innerhalb der Altersgruppen deutet auf heterogene soziolinguistische Muster hin. Systematischere Effekte hätten sich potenziell isolieren lassen, wenn die Teilnehmenden, analog zu den stimulusspezifischen Effekten, anhand eines erweiterten Spektrums an Hintergrundmerkmalen modelliert worden wären.

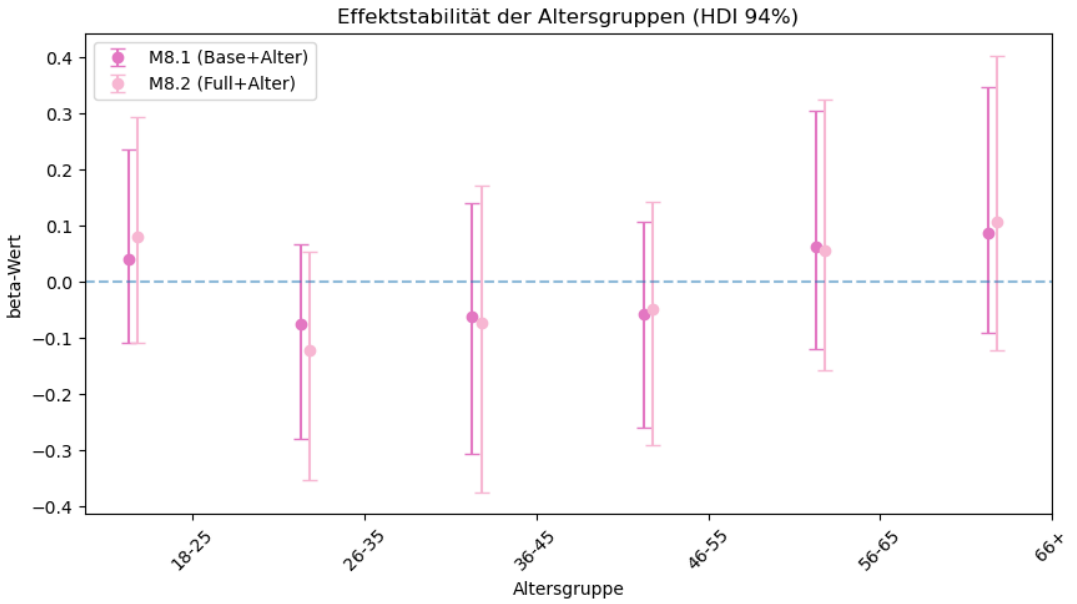

Abbildung 42: Effektstabilität der Altersgruppen (M8.1, M8.2). Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für den Alters-Effekt über zwei Modelle. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

6.4.2.4 Dialektexpositionseffekte

Abschliessend erweist sich die Dialektexposition als ungeeigneter Prädiktor für das Bewertungsverhalten. Der Koeffizient schrumpft gegen Null, sobald das Modell für die stimulus- und personenspezifischen Kerneffekte kontrolliert wird (*siehe Abbildung 43*). Folgerichtig verzeichnet das Modell M9.2 (*Full+Expo*) im Vergleich zu M4 (*Full*) keinerlei Zuwachs an prädiktiver Kraft.

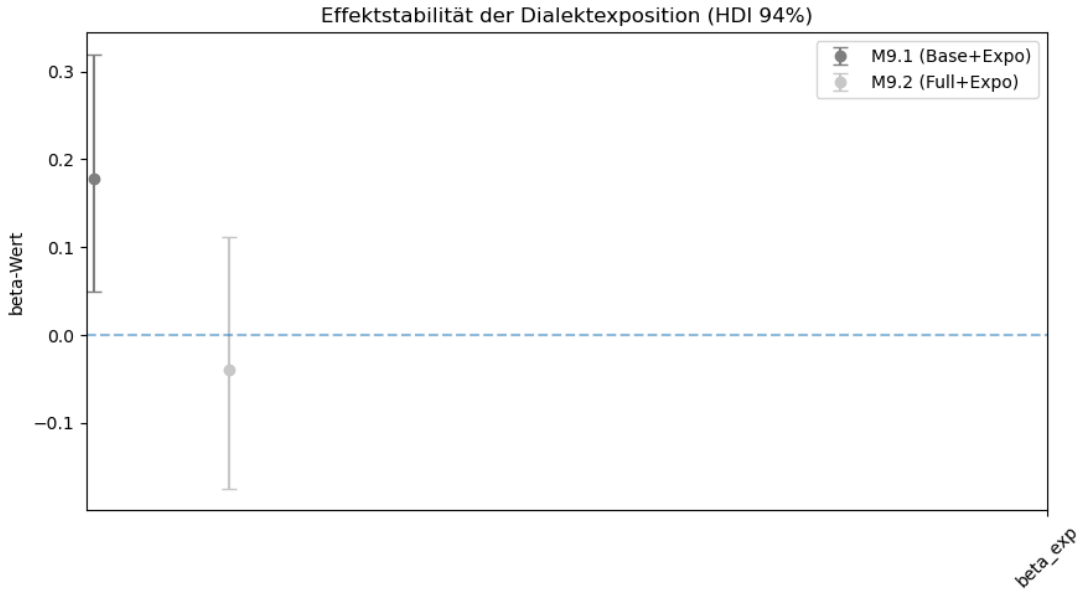

Abbildung 43: Effektstabilität der Dialektexposition (M9.1, M9.2). Geschätzte Regressionskoeffizienten (Beta-Werte, Y-Achse) für den Dialektexpositions-Effekt über zwei Modelle. Punkte markieren posteriore Mediane, vertikale Fehlerbalken das 94%-Highest-Density-Interval (HDI), die horizontal gestrichelte Linie den Nullpunkt (kein Effekt). Farben codieren die Modelle.

7 Diskussion

Das primäre Ziel dieser Arbeit bestand in der Entwicklung einer robusten Evaluationsmethodik, um die inhärent subjektive Wahrnehmung von Dialektauthentizität in schweizerdeutschen TTS-Systemen empirisch quantifizierbar zu machen. Durch die systematische Verknüpfung eines gezielten Forschungsdesigns, der Eigenentwicklung einer webbasierten Annotationsplattform sowie der Anwendung probabilistischer Auswertungsverfahren (*Bayessche ordinale logistische Regression*) konnte dieses Vorhaben erfolgreich umgesetzt werden. Die nachfolgende Reflexion ordnet die daraus gewonnenen zentralen Befunde kritisch ein und diskutiert deren Implikationen für die Beurteilung synthetischer Sprache.

7.1 Reflexion der Ergebnisse

Dominanz der Subjektivität: Ein zentrales Ergebnis dieser Arbeit ist die deutliche Dominanz der Subjektivität bei der Evaluation synthetischer schweizerdeutscher Sprache. Die Varianzanalysen der bayesschen Modelle bestätigen, dass der personenspezifische Bias sowie die Annotationssicherheit die Bewertungsunterschiede massgeblich determinieren. Die Wahrnehmung von Dialektauthentizität folgt somit keinem absoluten absoluten, objektiven Massstab, sondern hängt stark von der individuellen Prägung und dem Vertrauen in das eigenes Urteil ab.

Stimulusspezifische Einflüsse: Bei der Betrachtung der stimulusspezifischen Parameter zeigt sich: Die linguistische und inhaltliche Struktur des Satzes hat einen signifikant stärkeren Einfluss auf die Bewertung als der Dialekt oder die Referenzstimme. Dies legt nahe, dass bestimmte phonetische oder syntaktische Konstrukte vom zugrunde liegenden TTS-Modell unterschiedlich gut verarbeitet und synthetisiert werden. Dadurch variiert die wahrgenommene Authentizität unabhängig von regionalen Klangfärbungen stark.

Methodische Befunde zur Aggregation: Ein wesentlicher Befund betrifft die Auswertung der Annotationsdaten. Direkte Aggregationsmethoden wie der Majority Vote erweisen sich für die Qualitätsbewertung von Dialektsynthesen als ungeeignet, da sie bei derart subjektiven Aufgabenstellungen aufgrund der tiefen IRR scheitern. Um personenspezifische Wahrnehmungsunterschiede, Bias-Faktoren und die ordinale Natur der Daten korrekt zu modellieren, ist der Einsatz probabilistischer Verfahren unerlässlich. Erst durch die Anwendung der bayesschen ordinalen logistischen Regression konnte diese vielschichtige Subjektivität statistisch fundiert abgebildet werden.

Kalibrierung und Selbsteinschätzung: Die Rolle der Kalibrierungsphase muss kritisch reflektiert werden. Als Instrument zur Gewichtung oder Validierung der Bewertungen liefert die empirische Kalibrierung keinen nachweisbaren Mehrwert im statistischen Modell. Im Gegensatz dazu zeigte die Selbsteinschätzung einen messbaren Effekt. Interessanterweise stimmte diese Einschätzung – mit Ausnahme der Dialektregion Graubünden – weitgehend mit effektiven Kalibrierungsresultaten überein. Dies deutet darauf hin, dass die Teilnehmenden ihre Fähigkeiten zur Dialektidentifikation präzise einschätzen konnten.

Ausblick zur dynamischen Steuerung: Der Ansatz, Kalibrierungsresultate zur dynamischen Steuerung der Annotationsreihenfolge einzusetzen, besitzt dennoch konzeptionelles Potenzial für künftige Studien. Ein solcher Mechanismus könnte die Anzahl benötigter Annotationen reduzieren und die Effizienz des Annotationsprozesses erhöhen. Ob dies in der Praxis zu validen Resultaten führt, lässt sich auf Grundlage der vorliegenden Daten jedoch nicht abschliessend beurteilen. Dies liegt unter anderem daran, dass die Stichprobe stark von Teilnehmenden aus

dem Zürcher Dialektraum dominiert wurde, was die Übertragbarkeit der Ergebnisse auf andere Regionen einschränkt.

7.2 Limitationen

Demografische Asymmetrie: Bei der Interpretation der Ergebnisse sind methodische und strukturelle Limitationen zu berücksichtigen, welche die Generalisierung der Ergebnisse einschränken. Eine massgebliche Einschränkung betrifft die demografische Zusammensetzung der Stichprobe. Das Feld der Teilnehmenden wird stark von der Dialektregion Zürich dominiert. Diese Asymmetrie erforderte komplexe statistische Kontrollmechanismen und erschwert allgemeingültige Aussagen für den gesamten deutschschweizerischen Sprachraum, insbesondere für geografisch unterrepräsentierten Regionen.

Diskrepanz zwischen Identifikation und Bewertung: Ein konzeptioneller Aspekt betrifft die Diskrepanz zwischen der reinen Identifikation und der qualitativen Bewertung von Dialekten. Zwar liess sich empirisch nachweisen, dass die Selbsteinschätzung der Dialekterkennungskompetenz einen messbaren Effekt auf das Bewertungsverhalten ausübt. Die Fähigkeit, einen Dialekt korrekt zu identifizieren, impliziert jedoch nicht zwangsläufig die Kompetenz, dessen Dialektauthentizität qualitativ sinnvoll zu bewerten. Diese Unterscheidung konnte im vorliegenden Studiendesign nicht vollständig aufgelöst werden.

Spezifität des TTS-Modells: Die Reichweite der Erkenntnisse wird durch die Spezifität des verwendeten TTS-Modells begrenzt. Alle evaluierten Stimuli basieren auf der Synthese einer einzigen Modellarchitektur. Aus den erhobenen Daten lässt sich keine Aussage ableiten, ob dieses Modell absolut betrachtet qualitativ gut oder schlecht ist. Da ein objektiver Referenzwert von Dialektauthentizität fehlt, zielte die statistische Auswertung primär darauf ab, die Einflussfaktoren auf das menschliche Bewertungsverhalten zu quantifizieren. Eine Einstufung der tatsächlichen Modellleistung war unter diesen Voraussetzungen nicht möglich.

Design der Kalibrierungsphase: Das experimentelle Design der Kalibrierungsphase unterliegt ebenfalls Einschränkungen. Da die Kalibrierungsstimuli in einer fixen, nicht randomisierten Reihenfolge präsentiert wurden, liessen sich Eingewöhnungseffekte statistisch nicht isolieren. Hinzu kommt, dass die Annotationsplattform retrospektive Korrekturen der abgegebenen Antworten zuliess. Ein möglicher Lernprozess während der Durchführung könnte somit zu nachträglichen Anpassungen durch die Annotierenden geführt haben, was eine präzise Messung des kontinuierlichen Lernfortschritts verunmöglichte.

7.3 Ausblick

Aus den Ergebnissen und Limitationen dieser Arbeit ergeben sich konkrete Ansätze für künftige Projekte:

Da der Satzeffekt einen dominanten Einfluss auf die Bewertungen aufweist, ist eine detailliertere linguistische Analyse der verwendeten Sätze sinnvoll. Es sollte untersucht werden, welche Lautkombinationen oder syntaktische Strukturen das XTTS-Modell bereits adäquat synthetisiert und wo systematische Fehler liegen, um das Modelltraining gezielt zu optimieren.

Für künftige Annotationsplattformen wird eine strikte Randomisierung der Stimulireihenfolge in der Kalibrierungsphase empfohlen, um Eingewöhnungs- oder Ermüdungseffekte auszuschliessen. Zudem muss die Plattform technisch so angepasst werden, dass nachträgliche

Korrekturen abgegebener Antworten nachvollziehbar erfasst werden. Nur so lässt sich ein möglicher Lernfortschritt der Teilnehmenden präzise messen.

Um die geografische Verzerrung auszugleichen, ist eine gezielte demografische Erweiterung der Stichprobe erforderlich. Künftige Erhebungen sollten vermehrt Personen aus den unterrepräsentierten Dialektregionen einbeziehen, um eine ausgewogenere Datenbasis zu schaffen.

Die gesammelten Annotationsdaten bieten eine fundierte Grundlage, um automatische Bewertungsmetriken wie UTMOS spezifisch für Schweizerdeutsch zu optimieren. Ein solches, auf menschlichen Bewertungen basierendes und entsprechend feinabgestimmtes Modell könnte die subjektive Wahrnehmung besser abbilden und die automatisierte Qualitätskontrolle künftiger Dialekt-TTS-Systeme massgeblich erleichtern.

8 Verzeichnisse

8.1 Literaturverzeichnis

[1] S. Chen *u. a.*, «VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers», 17. Juni 2024, *arXiv*: arXiv:2406.05370. doi: 10.48550/arXiv.2406.05370.

[2] H. Azzuni und A. E. Saddik, «Voice Cloning: Comprehensive Survey», 1. Mai 2025, *arXiv*: arXiv:2505.00579. doi: 10.48550/arXiv.2505.00579.

[3] S. Stucki, «No Shots Required: Zero-Shot Voice-Adaptation TTS for Swiss German Dialects», Masterarbeit, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2024. Zugegriffen: 20. Februar 2026. [Online]. Verfügbar unter: https://www.zhaw.ch/storage/engineering/institute-zentren/cai/studentische_arbeiten/Fall_2024/MSE_MT_25_ciel_Stucki_TTS_for_Swiss_German.pdf

[4] F. Rash, *The German Language in Switzerland: Multilingualism, Diglossia and Variation*. Bern: Peter Lang, 1998.

[5] Y. Scherrer, «Machine translation into multiple dialects: The example of Swiss German», in *7th SIDG Congress - Dialect 2.0*, Wien, 2012. Zugegriffen: 29. Mai 2026. [Online]. Verfügbar unter: https://scispace.com/pdf/machine-translation-into-multiple-dialects-the-example-of-4l8yir8o6q.pdf

[6] S. Stucki, J. Deriu, und M. Cieliebak, «Voice Adaptation for Swiss German», 28. Mai 2025, *arXiv*: arXiv:2505.22054. doi: 10.48550/arXiv.2505.22054.

[7] S. Stucki, M. Cieliebak, und J. Deriu, «SwissGPC v1.0 -- The Swiss German Podcasts Corpus», 24. September 2025, *arXiv*: arXiv:2509.19866. doi: 10.48550/arXiv.2509.19866.

[8] S. Stucki und P. Randjelovic, «Automatic Detection of Swiss German Dialects using Wav2Vec», Bachelorarbeit, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2023. Zugegriffen: 20. Februar 2026. [Online]. Verfügbar unter: https://www.zhaw.ch/storage/engineering/institute-zentren/cai/studentische_arbeiten/Spring_2023/Spring23_BA_ciel_Dialect_Recognition_Swiss_German_Schneider_Frei.pdf

[9] C. Frei und P. Schneider, «Automatic identification of Swiss German dialects using large language models», Bachelorarbeit, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2023. doi: 10.21256/zhaw-29667.

[10] L. Bolliger und S. Waldburger, «Automatische Erkennung schweizerdeutscher Dialekte anhand von Audiodaten via Phonemtranskriptionen», Bachelorarbeit, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, 2024. doi: 10.21256/zhaw-31985.

[11] C. Paonessa, Y. Schraner, J. Deriu, M. Hürlimann, M. Vogel, und M. Cieliebak, «Dialect Transfer for Swiss German Speech Translation», in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, und K. Bali, Hrsg., Singapore: Association for Computational Linguistics, Dez. 2023, S. 15240–15254. doi: 10.18653/v1/2023.findings-emnlp.1018.

[12] X. Tan, T. Qin, F. Soong, und T.-Y. Liu, «A Survey on Neural Speech Synthesis», arXiv.org. Zugegriffen: 29. Mai 2026. [Online]. Verfügbar unter: https://arxiv.org/abs/2106.15561v3

[13] Y. Jia *u. a.*, «Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis», arXiv.org. Zugegriffen: 29. Mai 2026. [Online]. Verfügbar unter: https://arxiv.org/abs/1806.04558v4

[14] E. Casanova *u. a.*, «XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model», arXiv.org. Zugegriffen: 29. Mai 2026. [Online]. Verfügbar unter: https://arxiv.org/abs/2406.04904v1
[15] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, und J. Yamagishi, «A review on subjective and objective evaluation of synthetic speech», *Acoustical Science and Technology*, Bd. 45, Nr. 4, S. 161–183, 2024, doi: 10.1250/ast.e24.12.
[16] International Telecommunication Union, *Methods for subjective determination of transmission quality*, ITU-T Recommendation P.800, Genf., 1996. Zugegriffen: 28. Mai 2026. [Online]. Verfügbar unter: https://www.itu.int/rec/t-rec-p.800-199608-i
[17] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, und H. Saruwatari, «UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022», 29. Juni 2022, *arXiv*: arXiv:2204.02152. doi: 10.48550/arXiv.2204.02152.
[18] K. Krippendorff, *Content analysis : an introduction to its methodology*, 2nd ed. Thousand Oaks (Calif.) : Sage Publications.
[19] N. Döring, *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin, Heidelberg: Springer, 2023. doi: 10.1007/978-3-662-64762-2.
[20] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, und D. B. Rubin, *Bayesian Data Analysis, Third Edition*. in Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. [Online]. Verfügbar unter: https://books.google.ch/books?id=ZXL6AQAAQBAJ
[21] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, und P.-C. Bürkner, «Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC», *Bayesian Anal.*, Bd. 16, Nr. 2, Juni 2021, doi: 10.1214/20-BA1221.
[22] «Thorsten-Voice Dataset 2022.10», Mozilla Data Collective. Zugegriffen: 8. Mai 2026. [Online]. Verfügbar unter: https://mozilladatacollective.com/datasets/cmm4b8y6100nomk07w678zb9d
[23] «Common Voice Scripted Speech 25.0 - German», Mozilla Data Collective. Zugegriffen: 8. Mai 2026. [Online]. Verfügbar unter: https://mozilladatacollective.com/datasets/cmn4rsdh6009unz07jdn2ol9p
[24] R. Jörg, *Untersuchungen zum Schwund des Präteritums im Schweizerdeutschen*, Bd. 52. in Basler Studien zur deutschen Sprache und Literatur, vol. 52. Bern: Francke, 1976.
[25] A. Leemann, C. Steiner, und M. Studerus, *Dialäktatlas : 1950 bis heute*. Zollikon, 2025. Zugegriffen: 5. März 2026. [Online]. Verfügbar unter: https://dialektatlas.ch/Dialaektatlas_2025_150dpi.pdf
[26] tarepan, *tarepan/SpeechMOS*. (10. Mai 2026). Python. Zugegriffen: 28. Mai 2026. [Online]. Verfügbar unter: https://github.com/tarepan/SpeechMOS
[27] «openai/whisper-large-v3 · Hugging Face». Zugegriffen: 28. Mai 2026. [Online]. Verfügbar unter: https://huggingface.co/openai/whisper-large-v3
[28] E. L. Dolev, C. F. Lutz, und N. Aepli, «Does Whisper understand Swiss German? An automatic, qualitative, and human evaluation», 9. Mai 2024, *arXiv*: arXiv:2404.19310. doi: 10.48550/arXiv.2404.19310.

8.2 Abbildungsverzeichnis

Abbildung 1: Schematische Darstellung einer Zero-Shot-TTS-Architektur mit separatem Speaker Encoder zur Extraktion von Stimmmerkmalen 4
Abbildung 2: Karte der schweizerdeutschen Dialektregionen mit Kantonszuweisungen 11
Abbildung 3: Selbsteinschätzungs-Matrix (Dialektidentifikation) 12
Abbildung 4: Kalibrierungsansicht mit Audioplayer, Dialektauswahl und Sicherheitsskala .. 13
Abbildung 5: Annotationsmaske mit Audioplayer, Transkription sowie den Bewertungsskalen für Authentizität und Sicherheit 15
Abbildung 6: UTMOS-Verteilung pro Modell 21
Abbildung 7: WER-Verteilung pro Modell 21
Abbildung 8: WER- und UTMOS-Verteilung nach Dialekt 23
Abbildung 9: WER- und UTMOS-Verteilung nach Referenzstimme 23
Abbildung 10: WER- und UTMOS-Verteilung nach Satz 24
Abbildung 11: Status der Teilnehmenden nach Studienphase 25
Abbildung 12: Verteilung der Annotationen pro teilnehmende Person 26
Abbildung 13: Verteilung der Teilnehmenden nach Altersgruppe 26
Abbildung 14: Verteilung der Teilnehmenden nach Dialektregion 27
Abbildung 15: Aufenthaltsorte der Teilnehmenden nach Kanton und Aufenthaltsdauer 27
Abbildung 16: Selbsteinschätzung (Dialektidentifikation) der Teilnehmenden 28
Abbildung 17: Kalibrierungsscore pro Dialekt 29
Abbildung 18: Zusammenhang Kalibrierungsscores und Selbsteinschätzung 29
Abbildung 19: Durchschnittliche Sicherheit vs. Korrektheit pro Kalibrierungsaufnahme 30
Abbildung 20: Zusammenhang zwischen durchschnittlicher Sicherheit und Korrektheit pro Kalibrierungsaufnahme 32
Abbildung 21: Durchschnittliche Korrektheit nach Position der Kalibrierungsaufnahme zur Analyse des Eingewöhnungseffekts 33
Abbildung 22: Verteilung der Annotationen nach Dialekt 34
Abbildung 23: Verteilung der Annotationssicherheit nach Dialekt 35
Abbildung 24: Varianzanteilskoeffizienten (VPC) 49
Abbildung 25: Varianzanteilskoeffizienten (VPC (nur mit stimulusspezifischen Effeken) 49
Abbildung 26: Effektstabilität der Dialekte 50
Abbildung 27: Posterior-Wahrscheinlichkeiten anhand Dialekt (M6.2) 51
Abbildung 28: Marginale Posterior-Wahrscheinlichkeiten anhand Dialekt (M6.2) 51
Abbildung 29: Effektstabilität der Referenzstimmen 52
Abbildung 30: Posterior-Wahrscheinlichkeiten anhand Referenzstimme (M6.2) 53
Abbildung 31: Marginale Posterior-Wahrscheinlichkeiten anhand Referenzstimme (M6.2).. 53
Abbildung 32: Kontrafaktische Posterior-Wahrscheinlichkeiten der Referenzstimmen nach Dialekt (M6.2) 54
Abbildung 33: Effektstabilität der Sätze 55
Abbildung 34: Marginale Posterior-Wahrscheinlichkeiten anhand Satz (M6.2) 56
Abbildung 35: Kontrafaktische Posterior-Wahrscheinlichkeiten der Sätze nach Dialekt (M6.2) 57
Abbildung 36: Verteilung der Beta-Werte für die Annotationssicherheit 58
Abbildung 37: Marginale Posterior-Wahrscheinlichkeiten anhand Annotationssicherheit (M6.2) 59
Abbildung 38: Effektstabilität Selbsteinschätzung (Dialektidentifikation) (M6.1, M6.2) 59
Abbildung 39: Marginale Posterior-Wahrscheinlichkeiten anhand Selbsteinschätzung (Dialektidentifikation) (M6.2) 60

Abbildung 40: Effektstabilität des Heimdialekts (M5, M6) 61
Abbildung 41: Effektstabilität des Kalibrierungsscores (M7.1, M7.2) 62
Abbildung 42: Effektstabilität der Altersgruppen (M8.1, M8.2) 63
Abbildung 43: Effektstabilität der Dialektexposition (M9.1, M9.2) 63

8.3 Tabellenverzeichnis

Tabelle 1: Modell-Ranking basierend auf UTMOS- und WER-Medianwerten. 20
Tabelle 2: Krippendorffs Alpha pro Majority Vote Variante. 38
Tabelle 3: Übersicht der bayesschen ordinalen logistischen Regressionsmodelle mit jeweiliger Basis und Erweiterung. 39
Tabelle 4: Kantone zu Dialektmapping. 45
Tabelle 5: Bayesscher Modellvergleich und Konvergenzdiagnostik der Regressionsmodelle 48

9 Anhang

9.1 Einsatz künstlicher Intelligenz (KI)

In verschiedenen Phasen dieser Arbeit kamen generative KI-Systeme und -Werkzeuge unterstützend zum Einsatz. Im Detail kamen folgende Tools zur Anwendung:

- **Programmierung und Implementierung:** Für die Codierung der Annotationsplattform, der zugehörigen Pipelines und der Auswertung wurden GitHub Copilot (mit automatischer Modellauswahl) innerhalb von VS-Code und Claude Sonnet 4.6 als Programmierassistenten eingesetzt.
- **Datengenerierung:** Für die Generierung der Sätze für die TTS-Synthese wurde das Modell Claude Sonnet 4.5 als Hilfsmittel verwendet.
- **Strukturierung und Review:** Gemini 3.1 sowie ChatGPT (Free Tier) dienten als Schreibassistenten zur Strukturierung des Berichts und als Peer-Review-Werkzeuge, um die technische Präzision und Klarheit der Inhalte zu überprüfen.
- **Sprachlicher Feinschliff:** Für die linguistische Überarbeitung und zur Korrektur der Formulierungen des finalen Textes wurde DeepL herangezogen.

Sämtliche durch diese Hilfsmittel generierten Ergebnisse, Konzepte und Code-Bestandteile wurden von der Autorenschaft kritisch geprüft, bei Bedarf angepasst und liegen in deren alleiniger Verantwortung.

9.2 Quellcode und technische Dokumentation

- **Annotationsplatform**: https://github.zhaw.ch/ba-tts-human-verification/tts-annotation-plattform
- **TTS-Pipeline**: https://github.zhaw.ch/ba-tts-human-verification/tts-pipeline
- **Audio/Voice-Cleanup-Skripts**: https://github.zhaw.ch/ba-tts-human-verification/voice-dataset
- **Auswertungsmodelle/Skripts/Plots**: https://github.zhaw.ch/ba-tts-human-verification/evaluation