Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

CAI Centre for Artificial Intelligence

Projektarbeit (Informatik)

Automatische Erkennung von Dialekten (Schweizerdeutsch und Englisch)

Autoren	Laura Bolliger Safiyya Waldburger
Hauptbetreuung	Dr. Jasmina Bogojeska Prof. Dr. Mark Cieliebak
Datum	22.12.2023

Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

Erklärung betreffend das selbstständige Verfassen einer Projektarbeit an der School of Engineering

Mit der Abgabe dieser Projektarbeit versichert der/die Studierende, dass er/sie die Arbeit selbständig und ohne fremde Hilfe verfasst hat. (Bei Gruppenarbeiten gelten die Leistungen der übrigen Gruppenmitglieder nicht als fremde Hilfe.)
Der/die unterzeichnende Studierende erklärt, dass alle zitierten Quellen (auch Internetseiten) im Text oder Anhang korrekt nachgewiesen sind, d.h. dass die Projektarbeit keine Plagiate enthält, also keine Teile, die teilweise oder vollständig aus einem fremden Text oder einer fremden Arbeit unter Vorgabe der eigenen Urheberschaft bzw. ohne Quellenangabe übernommen worden sind.

Bei Verfehlungen aller Art treten die Paragraphen 39 und 40 (Unredlichkeit und Verfahren bei Unredlichkeit) der ZHAW Prüfungsordnung sowie die Bestimmungen der Disziplinarmassnahmen der Hochschulordnung in Kraft.

Ort, Datum:	**Name Studierende:**
Zürich, 22.12.2023	Safiyya Waldburger 
Zürich, 22.12.2023	Laura Bolliger 

Abstract

For multilingual systems that need to understand language, the differentiation of languages and dialects is an important component. For low-resource languages such as Swiss German, which is spoken by comparatively few people, this is a greater challenge. Especially when it comes to spoken language, which is also characterized by phonemes, intonation and speaker-specific characteristics. This paper deals with a classification model based on the pretrained model Whisper, which was trained with data sets from the recent Swiss corpora SDS-200 and STT4SG-350 as well as comparatively with an English corpus and categorizes short audio samples with spoken language into one of seven or six dialect regions. For this purpose, the model architecture of a previous study was adopted and its results were reproduced. Building on this, further experiments were carried out with the same model, which explored the various possibilities of compiling training data and the difficulties of dialect recognition. The experiments showed that a larger number of speakers and a balanced number of samples per dialect region and speaker are favorable, but that the results deteriorate if the number of samples per speaker is too large. In addition, the results of the training with the English corpus suggest that English has a certain advantage over Swiss German thanks to the pre-training of Whisper on more English data. The fact that samples of individual speakers probably do not only have characteristics of one dialect, but could potentially be characterized by several regions, proved to be particularly difficult in terms of dialect recognition. These findings can be used for further work in order to achieve even better results with an optimized data set and model and to deepen interesting aspects of dialect recognition.

Zusammenfassung

Für multilinguale Systeme, die Sprache verstehen müssen, ist die Unterscheidung von Sprachen und Dialekten ein wichtiger Bestandteil. Für Low-Resource-Sprachen wie Schweizerdeutsch, die von vergleichsweise wenigen Menschen gesprochen werden, ist dies eine grössere Herausforderung. Erst recht, wenn es um gesprochene Sprache geht, die zusätzlich von Phonemen, Betonungen, aber auch sprecher:innenspezifischen Eigenschaften geprägt sind. Diese Arbeit beschäftig sich mit einem Klassifizierungsmodell basierend auf dem Pretrained-Model Whisper, das mit Datensets aus den neueren Schweizer Korpora SDS-200 und STT4SG-350 sowie vergleichend mit einem englischen Korpus trainiert wurde und kurze Audiosamples mit gesprochener Sprache in eine von sieben beziehungsweise sechs Dialektregionen einteilt. Dafür wurde die Modell-Architektur einer Vorgängerarbeit übernommen und zunächst deren Ergebnisse reproduziert. Darauf aufbauend wurden mit dem gleichen Modell weitere Experimente durchgeführt, die einerseits die verschiedenen Möglichkeiten, Trainingsdaten zusammenzustellen und die Schwierigkeiten der Dialekterkennung erkundeten. Die Experimente zeigten,

dass eine grössere Anzahl Sprecher:innen und eine ausgeglichene Anzahl Samples pro Dialektregion und Sprecher:in vorteilhaft sind, dass sich aber die Ergebnisse ab einer zu grossen Anzahl Samples pro Sprecher:in verschlechtern. Zudem lassen die Resultate des Trainings mit dem englischen Korpus vermuten, dass Englisch dank dem Pretraining von Whisper auf mehr englischen Daten einen gewissen Vorteil gegenüber Schweizerdeutsch hat. Als besonders schwierig bei der Dialekterkennung stellte sich heraus, dass Samples einzelner Sprecher:innen wahrscheinlich nicht nur Merkmale eines Dialektes aufweisen, sondern potenziell von mehreren Regionen geprägt sein könnten. Diese Erkenntnisse können für weitere Arbeiten genutzt werden, um noch bessere Resultate mit einem optimierten Datenset und Modell zu erzielen und interessante Aspekte der Dialekterkennung zu vertiefen.

Inhaltsverzeichnis

1. Einleitung

Die Dialekterkennung gehört zur automatischen Spracherkennung (ASR), die sich mit der Verarbeitung und Umwandlung gesprochener Sprache befasst und ein Teilgebiet des Natural-Language-Processings (NLP) ist [1]. Viele alltägliche Anwendungen wie die Sprachassistentin Siri basieren auf automatischer Erkennung von Sprache [2].

Die Herausforderung in der maschinellen Spracherkennung liegt in der Komplexität und Vielfalt der Sprache als Gegenstand wie beispielsweise in der Mehrdeutigkeit von Wörtern im Kontext oder in den phonologischen Unterschieden, die durch Dialekte oder Akzente gegeben sind [1].

Insbesondere die Erkennung von sogenannten Low-Resource-Sprachen wie des Schweizerdeutschen ist anspruchsvoll [3]. Diese Sprachen werden von vergleichsweise wenigen Menschen gesprochen und bisherige Modelle im Bereich Deep-Learning benötigten meist grosse Mengen an Daten für die Merkmalserkennung von Sprachen [3]. Mit den in den letzten Jahren erschienenen Pretrained-Large-Language-Models, die Gelerntes auf andere Sprachen übertragen können und so auch mit einer begrenzten Menge an Daten auskommen, eröffnen sich neue Möglichkeiten für die Erkennung von Low-Resource-Sprachen [4][5].

Diese Arbeit konzentriert sich auf die Erkennung und Klassifizierung von gesprochenen schweizerdeutschen Dialekten, wobei die Begriffe «Dialekterkennung» und «Dialektidentifizierung» austauschbar verwendet werden. Aufbauend auf einer Vorgängerarbeit und den in den zwei letzten Jahren zugänglich gemachten Schweizer Korpora SDS-200 und STT4SG-350 werden Experimente zur Dialekterkennung durchgeführt und Einflussfaktoren auf die Erkennung von schweizerdeutschen Dialekten mithilfe von Pretrained-Models erkundet. Ergänzend wird ein englischer Korpus hinzugezogen, um die Leistung der Modelle bei einer High-Resource-Sprache zu vergleichen.

Das erste Kapitel befasst sich mit dem grundlegenden Hintergrundwissen, das für die nachfolgenden Experimente relevant ist. Im zweiten Kapitel werden die Ergebnisse der Vorgänger reproduziert und die darauf aufbauenden Experimente erläutert. Zum Abschluss erfolgt eine Diskussion der wesentlichen Erkenntnisse, begleitet von einem Ausblick.

2. Theorie und Hintergrund

Für die Experimente zur automatischen Dialekterkennung bei gesprochener Sprache sollen in dieser Arbeit die neusten Werkzeuge des Deep-Learnings für Natural-Language-Processing genutzt werden. Insbesondere Large-Language-Models (LLM), die mit einer grossen Menge von Daten trainiert wurden, bieten ein grosses Potenzial für die Dialekterkennung. In diesem Kapitel wird das Hintergrundwissen erläutert, das für das Verständnis der Experimente basierend auf solchen LLMs von Bedeutung ist.

2.1 Grundbausteine für die Modelle

2.1.1 Pretrained Models

Ein Pretrained-Model wurde auf umfangreichen und vielfältigen Datensätzen trainiert und ist dadurch bereits in der Lage Repräsentationen dieser Daten zu generieren [6]. Als solches kann es als leistungsstarker Baustein eines weiteren Modells[1] genutzt werden. Wird die Instanz eines Pretrained-Models für eine bestimmte Aufgabe weitertrainiert, so ähnelt es einem Marathonläufer mit Vorsprung, der nicht mehr von der Startlinie aus beginnen muss, sondern bereits ein gewisses Mass an Wissen mit sich bringt. Während also beim Training eines Modells ohne Pretraining zufällige Gewichten initialisiert werden, können die Gewichte des Pretrained-Models übernommen werden [6][7]. Diese Gewichte werden dann mit Finetuning angepasst, damit das Modell eine spezifischere, aber ähnliche Aufgabe bewältigen kann, wie für die es bereits trainiert wurde [6]. Wie in der Einleitung angedeutet, bieten Pretrained-Models, die auf verschiedenen Sprachen vortrainiert wurden und ihre sprachübergreifend gelernten Repräsentationen transferieren können, Chancen für die Erkennung von Low-Resource-Sprachen [7].

2.1.2 Transformer-Modelle

Transformer basieren auf einer Deep-Learning-Architektur, die es Maschinen ermöglicht, natürlichsprachliche Texte zu verstehen und zu generieren [8][9]. Sie können komplexe Beziehungen zwischen Daten erfassen, wodurch sie für Aufgaben im Bereich des Natural-Language-Processings (NLP) wie beispielsweise Übersetzungen oder Textgenerierungen geeignet sind[2]. Die Architektur besteht aus zwei Hauptkomponenten: dem Encoder und dem Decoder. Im Folgenden wird vor allem erstere Komponente näher betrachtet, da nur diese für die angewandten Modelle in dieser Arbeit von Belang ist.

Der Encoder (links in der Abbildung 1) wandelt Eingabedaten, in diesem Fall Wörter oder Tokens, in numerische Repräsentationen um [8][9][11]. Jedes Wort oder Token wird zu einem hochdimensionalen Vektor, genannt Word-Embedding, umgewandelt (siehe Input Embedding in Abbildung 1)[9]. Diese Word-Embeddings sind im Vektorraum so angeordnet, dass semantisch ähnliche Wörter nahe beieinander liegen [9]. Da die Wörter parallel in den Encoder eingegeben werden, muss ihre Reihenfolge im Satz

[1] Die englischen Begriffe «Model», Plural «Models», und ihre entsprechende deutsche Übersetzung «Modell» respektive «Modelle» werden in dieser Arbeit gleichbedeutend eingesetzt.

[2] Die Transformer haben sich im Bereich NLP dank des Self-Attention-Mechanismus durchgesetzt, der die Beziehung zwischen den Wörtern erfassen kann, sowie dank ihrer Performanz, indem sie parallele Verarbeitung von Eingaben erlauben [8]-[10].

berücksichtigt werden [10]. Dazu werden Positional Encodings verwendet, die die gleiche Dimensionalität wie die ursprünglichen Word-Embeddings aufweisen und zu diesen addiert werden [8][11].

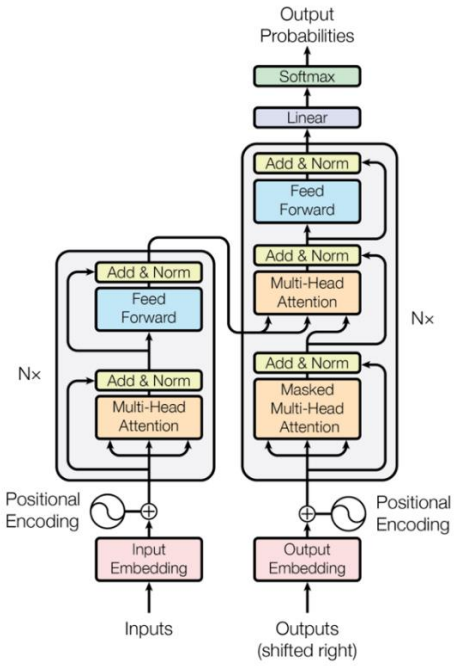


Abbildung 1: Transformer-Architektur: Encoder links, Decoder rechts [11]

Nachdem nun für jedes der eingegebenen Wörter eine Vektorrepräsentation generiert wurde, die die Semantik des einzelnen Wortes sowie dessen Position erfasst, wird mithilfe eines Self-Attention-Mechanismus die Semantik jedes Wortes im Kontext des eingegebenen Satzes errechnet (siehe Abbildung 1 «Multi-Head Attention»). Multi-Head-Attention kann als eine Art mehrfach durchgeführten Self-Attention-Prozess verstanden werden [9]. Je relevanter ein Wort in demselben Satz für die kontextuelle Bedeutung des betrachteten Wortes ist, desto höher der Wert [9]. Nun enthält die bisherige Vektorrepräsentation jedes Wortes ihre kontextuelle Bedeutung. Die Vektorrepräsentation wird anschliessend in einem Feed-Forward-Netzwerk verarbeitet, wodurch der Transformer vertieftere Repräsentationen lernt [9].

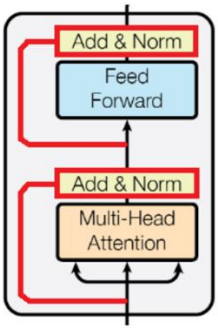


Abbildung 2: Teil des Transformer-Encoders aus Abbildung 1 [9]

Die Abzweigungen in Abbildung 2 die die Multi-Head-Attention- respektive Feed-Forward-Netzwerk-Schicht umgehen, dienen dazu, dass Informationen unverarbeitet in die oberen Schichten weitergereicht werden [9]. Dies optimiert den Lernvorgang zusätzlich [9]. Die Add- und die Norm-Schicht verbinden schliesslich die unverarbeiteten und verarbeiteten Resultate [9]. Damit hat der Encoder eine abstrakte Darstellung generiert, die Eigenschaften und Merkmale der Eingabe einfängt. Diese abstrakte Darstellung wird als latente Repräsentation bezeichnet [9].

Der Decoder erzeugt aus diesen latenten Repräsentationen Text für eine bestimmte Aufgabe [11]. Wie erwähnt, wird der Decoder für diese Arbeit jedoch nicht verwendet, da es sich bei der Dialekterkennung um eine Klassifikationsaufgabe handelt und somit ausschliesslich die generierten latenten Repräsentationen von Interesse sind. Mithilfe dieser Repräsentationen kann gesprochene Sprache analysiert und klassifiziert werden.

In den folgenden beiden Abschnitten werden sowohl das wav2vec- als auch das Whisper-Modell vorgestellt, die auf der Transformer-Architektur aufbauen. Beide Pretrained-Models verarbeiten Sprachsignale mithilfe von Encodern, um latente Repräsentationen zu erstellen.

2.1.3 wav2vec

wav2vec 2.0[3] wurde speziell für ein Finetuning auf einzelne Sprachen mit gelabelten Daten designt [12]. Im Gegensatz zu anderen Transformer-Modellen ist seine Architektur auf ein Pretraining mit ungelabelten Daten ausgelegt [12]. Dies hat den Vorteil, dass potenziell mehr Daten genutzt werden können, da wav2vec nicht auf aufwendig hinzugefügt Labels angewiesen ist. Abbildung 3 stellt den groben Aufbau des Modells dar. Die Audiodaten werden in roher Wellenform X in ein Encoder-Netzwerk eingespeist, das eine latente Repräsentationen Z der Audiosamples erstellt [12]. Diese werden dann teils maskiert, dem Transformer übergeben und zu Kontextrepräsentationen C umgewandelt [12]. Andererseits werden diese unmaskiert in eine quantisierte Repräsentation Q umgewandelt [12]. Das heisst, die Werte der Repräsentation werden aus einem kontinuierlichen Raum auf einen diskreten Raum abgebildet. Die Quantisierung ist jedoch für das Konzept des Modells an sich nicht entscheidend. Wichtiger ist, dass für das Training so je maskierte und unmaskierte Versionen der Repräsentationen vorhanden sind. Beim Training vergleicht das Modell die Kontextrepräsentationen und quantisierten Repräsentationen, indem es einen Loss zwischen den zwei Repräsentationen berechnet [12]. Ziel ist, diesen Loss zu minimieren und die Kontextrepräsentationen den passenden quantisierten Repräsentationen anzunähern und von den nicht passenden quantisierten Repräsentationen klarer zu unterscheiden [12]. Dieser Vorgang wird Contrastive-Learning genannt [12].

Für die Arbeit wird die multilinguale Form Wav2Vec2-XLS-R-300M verwendet, deren Architektur auf wav2vec 2.0 basiert und die mit 436000 Stunden Audiodaten und 128 Sprachen trainiert wurde und 300 Millionen Parameter beinhaltet [14]. Eine der 128 Sprachen ist Deutsch mit einem relativ grossen Datenanteil [14]. Schweizerdeutsch wird jedoch nicht aufgezählt [14].

[3] Die ursprüngliche Form des wav2vec-Modells ist archetektonisch etwas anders aufgebaut, verfolgt jedoch das gleiche Ziel wie sein verbessertes Nachfolgermodell [13].

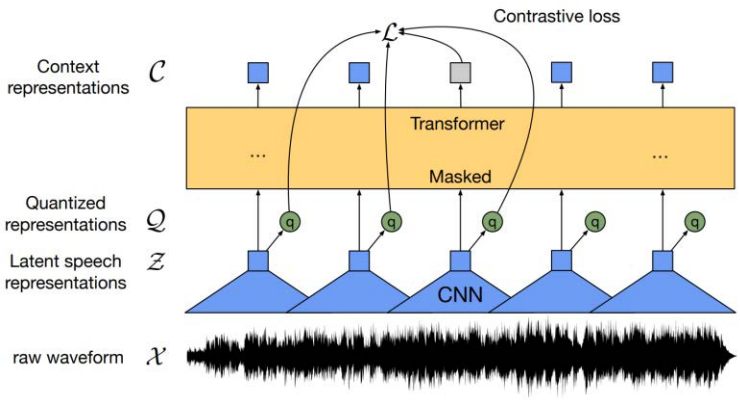


Abbildung 3: Architektur von wav2vec 2.0 [12]

2.1.4 Whisper

Das Whisper-Modell entstand aus der Motivation heraus, die Herausforderungen respektive Limitierung bisheriger ASR (Automatic Speech Recognition) Modelle wie beispielsweise wav2vec, die Repräsentationen aus ungelabelten Daten lernen, zu überwinden [15]. Den bisherigen Modellen fehlte es an effektiven Decodern und das erforderliche und herausfordernde Finetuning erschwerte ihren breiten Einsatz erheblich [15]. Zudem zeigen diese Modelle zwar eine starke Mustererkennung innerhalb ihres Trainingsdatensatzes, jedoch fällt es ihnen schwer, diese Muster auf andere Datensätze zu übertragen [15].

Das Whisper-Modell wurde mit 680'000 Stunden gelabelten, mehrsprachigen Audiodaten trainiert, wobei ein Drittel dieser Daten aus 96 nicht-englischen Sprachen unter anderem auch Deutsch stammt [15][16]. Schweizerdeutsch gehört jedoch nicht zu diesen Sprachen. Die Trainingsdaten umfassen verschiedene Quellen über mehrere Domänen und Datensätze hinweg, um die Generalisierbarkeit des Modells zu verbessern [15][17]. Darüber hinaus soll Whisper verglichen mit älteren ASR-Modellen robuster gegen Rauschen und Hintergrundgeräusche sein [15].

Das Modell verarbeitet 30-sekündige Audioabschnitte, die auf 1600 Hz abgetastet wurden [15][17]. In einem ersten Schritt werden diese Abschnitte in sogenannte Log-Mel-Spektrogramm-Repräsentationen umgewandelt [15][17]. Anschliessend werden diese in der Encoder-Decoder-Transformer-Architektur verarbeitet, wie in Abschnitt 2.1.2 beschrieben [17]. Am gleichen Audiosignal können verschiedene Aufgaben wie beispielsweise Transkription, Übersetzung, Erkennung von Sprachaktivitäten und Identifikation von Sprachen durchgeführt werden [17]. Wie bereits in Abschnitt 2.1.2 erwähnt, wird der Encoder von Whisper in dieser Arbeit für die Klassifikation von Dialekten verwendet.

Whisper ist in fünf verschiedenen Grössen mit unterschiedlicher Anzahl Parameter erhältlich. Diese sind in der Tabelle 1 ersichtlich [15].

MODEL	LAYERS[4]	PARAMETERS
TINY	4	39M
BASE	6	74M
SMALL	12	244M
MEDIUM	24	769M
LARGE	32	1550M

Tabelle 1: Whisper Modelle in unterschiedlichen Grössen

2.2 Datengrundlage

Grundlegend für das Finetuning der vorgestellten Pretrained-Models sind qualitativ hochwertige, gelabelte Audiodaten. Besonders bei Schweizerdeutsch ist es jedoch nicht einfach, genug Audiodaten zu beschaffen. So gibt es dialektspezifische Korpora wie der Swiss Parliaments Corpus (SPS) für das Berndeutsche, die für die Dialekterkennung weniger gut verwendbar sind [18]. Der diversere ArchiMob-Korpus ist zwar auf eine möglichst grosse Dialektvarietät ausgelegt, enthält jedoch sehr lange Samples und nur 555 Aufnahmen von wenigen Sprecher:innen [19]. Wobei für die Klassifikation eine möglichst grosse Anzahl Samples und Sprecher:innen praktischer sind. Besser eignen sich die beiden Korpora SDS-200 und STT4SG-350, die deutlich mehr Samples und Sprecher:innen beinhalten [20][21]. Um die Modelle auf ihr allgemeines Potenzial für die Dialekterkennung untersuchen zu können und mit dem Schweizerdeutschen vergleichen zu können, soll zusätzlich ein englischer Korpus eingesetzt werden.
Für das Englische ist die Auswahl an Korpora grösser, doch um einen guten Vergleich für das Schweizerdeutsche zu haben, sind besonders britische Dialekte interessant, die auch geographisch nahe beieinander auftreten. Als ähnlich aufgebauten Korpus wie der SDS-200- und der STT4SG-350-Korpus bot sich ein frei zugänglicher Korpus von Google an [22]. Die drei Korpora werden in den folgenden Abschnitten nochmals einzeln vorgestellt.

2.2.1 SDS-200-Korpus

Der SDS-200-Korpus enthält rund 200 Stunden Audiodaten à 152251 Audioclips von nicht ganz 3816 Sprecher:innen aus der deutschsprachigen Schweiz [20]. Ein Clip ist durchschnittlich 4.8 Sekunden lang [20]. Die Audio-Samples wurden mithilfe von Crowdsourcing gesammelt [20]. Dabei wurde ein Online-Tool erstellt, mit dem Sprecher:innen einen hochdeutschen Satz aus einer Zeitung oder aus dem German-Common-Voice-Korpus erhielten und in ihren Dialekt übersetzen und aufnehmen mussten [20]. Dazu wurden Herkunftskanton, Postleitzahl des Herkunftsorts, Alter und Geschlecht der Sprecher:innen erfasst [20]. Die Sprecher:innen wurden über Medien angeworben und mit zwei verschiedenen Wettbewerben dazu animiert mitzumachen und möglichst viele Samples aufzunehmen [20]. Zudem wurden die Sprecher:innen gebeten, ihre Samples gegenseitig zu validieren [20]. Um die Datenqualität zusätzlich abzusichern, wurden die Daten danach nochmals gefiltert [20]. In Abbildung 4 ist die Anzahl Samples pro Kanton, Geschlecht und Alter zu sehen. Es ist hervorzuheben, dass die Verteilungen der Samples bezüglich

[4] Layers für Anzahl Encoder respektive Decoder Schichten.

der Kanton, Geschlecht und Alter durch das Crowdsourcing nicht ausgeglichen sind, womit beim Trainieren eines Modells umgegangen werden muss.

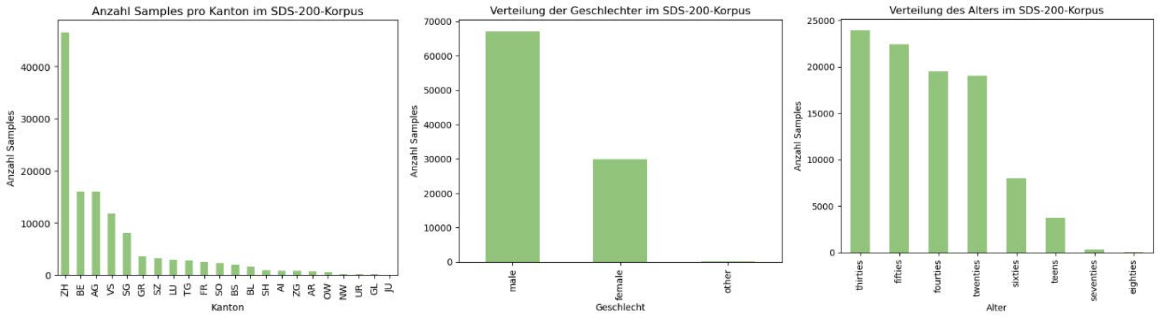


Abbildung 4: Statistiken zum SDS-200-Korpus

2.2.2 STT4SG-350-Korpus

Der STT4SG-350-Korpus ist nochmals deutlich grösser als der SDS-200-Korpus und bisher der grösste Schweizerdeutsch-Korpus [21]. Er enthält 343 Stunden à über 200000 Audiosamples von 316 Sprecher:innen [21]. Gesammelt wurden die Daten über das gleiche Verfahren wie beim SDS-200-Korpus, nur wurden die Audiodateien als verlustlose FLACs statt als MP3 gespeichert [21]. Allerdings wurden die Sprecher:innen gezielter rekrutiert [21]. Die Qualität der Samples wurde bei einem sinnvollen Anteil der Samples manuell überprüft [21]. Neben den Audiodaten mussten die Sprecher:innen Alter, Geschlecht und Postleitzahl des Herkunftsorts deklarieren und eine von sieben Dialektregionen für ihren Dialekt auswählen. Diese Regionen sind Zürich, Ostschweiz, Graubünden, Innerschweiz, Bern, Basel und Wallis [21].

Die gesammelten Daten wurden jeweils einem Trainings-, Validierungs- und Testsplit zugeordnet, die keine überschneidenden Sprecher:innen haben und in der Anzahl Sprecher:innen und Samples pro Dialektregion ausgeglichen sind [21]. Dazu wurde die Anzahl Samples pro Sprecher:in für das Testdatenset auf 368 und für das Trainingsdatenset auf 1112 limitiert und es wurden um die 45 Sprecher:innen pro Region rekrutiert [21]. Dies führt, wie in Abbildung 5 dargestellt, zu einer ausgeglichenen Anzahl Samples pro Dialektregion über alle Splits hinweg. Auch die Anzahl Samples pro Geschlecht ist nahezu gleich auf. Die Altersverteilung zeigt eine Mehrheit von jüngeren Sprecher:innen.

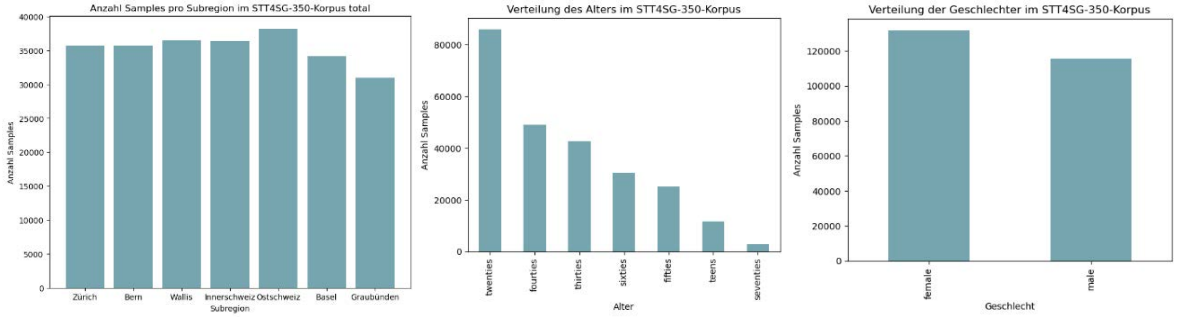


Abbildung 5: Statistiken zum STT4SG-350-Korpus

2.2.3 Englischer Korpus

Der in «Opensource Multispeaker Corpora of the English Accents in the British Isles» [22] beschriebene Korpus von Google ist mit 31 Stunden à 17877 Samples von 120 Sprecher:innen deutlich kleiner als der SDS-200- und der STT4SG-350-Korpus. Er enthält Aufnahmen von jeweils einem kurzen Satz von Sprecher:innen aus Grossbritannien mit sechs verschiedenen Akzenten beziehungsweise Dialekten, wobei sich die Sprecher:innen auch in diesem Korpus die Dialektregion selbst zugeordnet haben [23]. Auf der Karte in Abbildung 6 ist eingezeichnet, wo sich die im Korpus vorhandenen Regionen Schottland, Irland, Wales, Northern England, Midlands und Southern England ungefähr befinden [23]. Die Sprecher:innen wurden zum einen intern bei Google und zum andern bei der University of Cardiff rekrutiert. Ein besonderes Augenmerk lag auf der Auswahl der Sätze, die aufgenommen wurden [23]. Sie wurden aus Quellen wie Wikipedia oder «Alice im Wunderland» so selektiert, dass möglichst viele phonemische Phänomene ausgeglichen vertreten sind und für das Training eines Modells ein breites Spektrum an Dialektmerkmalen geboten wird [23]. Obwohl es sich im Vergleich zu anderen englischen Korpora um einen kleinen Korpus handelt, ist er von hoher Qualität und sorgfältig kuratiert.

Abbildung 6: Karte der Dialektregionen Englands [23]

Abbildung 7 zeigt die Anzahl Samples pro Dialektregion und Geschlecht. Wie beim SDS-200-Korpus sind die Verteilungen nicht ausgeglichen. Dies muss beim Training eines Modells vor allem bei der Anzahl Samples pro Dialektregion berücksichtigt werden. Die Anzahl Samples pro Sprecher:in wurde hingegen auf minimal 150 und maximal 300 beschränkt und ist folglich ausgeglichen [23].

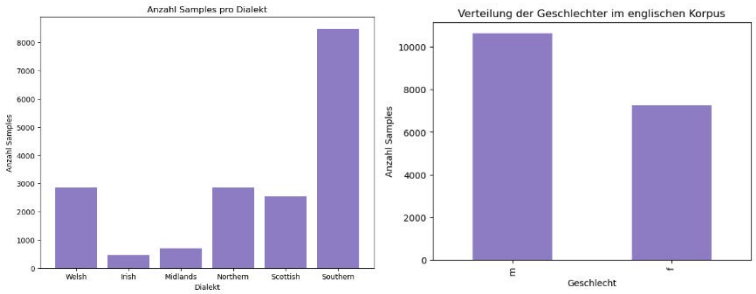


Abbildung 7: Statistiken zum englischen Korpus

2.3 Vorhergehende Arbeiten

Dieser Abschnitt bietet eine Übersicht über bisherige Forschung, die den Kontext für diese Arbeit bildet. Zunächst wird eine Auswahl interessanter Arbeiten zur Dialekterkennung bei verschiedenen Sprachen vorgestellt. Weil spezifisch zur Dialekterkennung beim Schweizerdeutschen weniger publiziert wurde, werden zusätzlich Arbeiten zu allgemeinen Spracherkennungsaufgaben einbezogen. Weiter werden die direkten Vorgängerarbeiten der ZHAW zusammengefasst, auf denen diese Arbeit aufbaut.

2.3.1 Allgemeine Forschung

In einer älteren Arbeit aus dem Jahr 2008 erreichten Mingliang und Yuguo eine Accuracy von 92.5 % bei der Erkennung von chinesischen Dialekten [24]. Sie setzten generative Modelle zur globalen Merkmalserkennung und diskriminative Modelle zur Klassifizierung der Dialekte ein [24]. Die Sprachsignale wurden mit Gaussian-Mixture-Models (GMM) in Tokens umgewandelt, mit N-Gramm analysiert und mit SVM und k-Means-Clustering klassifiziert [24].

Nour-Eddine und Abdelkader verwendeten im Jahr 2015 ebenfalls SVM in Kombination mit einem GMM-Model, um fünf arabische Maghreb-Dialekte zu identifizieren und erreichten dabei eine Precision von 80.49 % [25].

Bougrine und Abdelali befassten sich im Jahr 2018 mit der Erkennung algerischer Dialekte und verglichen die Leistung von Deep-Neural-Networks (DNN) mit SVMs. Beide Methoden zeigten sich ähnlich effektiv mit einer durchschnittlichen Precision von 47 % respektive 47,8 % [26]. Im Gegensatz zu den meisten Arbeiten, die sich mit der Erkennung arabischer Dialekte beschäftigen, wurden hier Dialekte innerhalb desselben geografischen Gebiets untersucht, also Dialekte, die näher beieinander liegen [26].

Im Jahr 2022 wurde eine Arbeit von Imaizumi et al. publiziert, die sich mit einem Multitasking-Lernansatz auseinandersetzte, bei dem Dialektidentifikation (DID) und automatische Spracherkennung für mehrere Dialekte (ASR) gleichzeitig in einem Modell trainiert werden [27]. Die Autoren stellen drei Modelle vor, die sie experimentell getestet haben: DID2ASR, ASR2DID und DID+ASR [27]. Die Modelle zeigten Verbesserungen in der Dialekterkennung und der ASR für japanische Dialekte im Vergleich zu konventionellen Ansätzen [27]. ASR2DID erwies sich als besonders wirksam für die DID-Aufgabe und

erreichte eine Accuracy von 86.5 % [27]. Demnach kann die gleichzeitige Berücksichtigung von ASR bei der Dialekterkennung für die Leistung des Modells vielversprechend sein.

Alrehaili et al. entwickelten im Jahr 2023 ein System zur Identifikation von acht arabischen Dialekten mittels Deep-Learning-Convolutional-Neural-Networks (CNN) [28]. In ihrer Arbeit verwendeten sie einen über die Regionen ausgeglichenen Datensatz mit 84 Samples pro Region [28]. Die Audioeingaben wurden zunächst in Spektrogramme umgewandelt und dann den CNNs für die Merkmalserkennung übergeben. Das System erzielte eine Accuracy von 83 % [28].

Im Rahmen der jährlichen Workshops der Varieties-and-Dialects-Kampagne (VarDial) wurde im Jahr 2018 eine Aufgabe zur schweizerdeutschen Dialekterkennung präsentiert [29] [30]. Die Aufgabe bestand darin, transkribierte Sätze aus vier Schweizer Regionen (BE, BS, LU und ZH) aus dem ArchiMob-Korpus zu identifizieren. Zusätzlich galt es, einen Überraschungsdialekt zu identifizieren, der erst im Testset auftauchte und nicht Teil des Trainingssatzes war [30]. Das Gewinnerteam Jauhiainen, Jauhiainen und Lindén erreichte einen Macro-F1-Score von 68.6 % [31]. Im darauffolgenden Jahr 2019 wurde dieselbe Aufgabe mit kleinen Anpassungen wie beispielsweise einer alternativen Form der Transkription präsentiert. Das beste Team erzielte einen Macro-F1-Score von 75.93 % [32].

Sicard, Gillioz und Pyszkowski haben sich in diesem Jahr 2023 mit der Transkription von schweizerdeutschen Dialekten zu Hochdeutsch beschäftigt [33]. Sie evaluierten die Leistung von Whisper und Wav2Vec2 XLS-R 1B unter dem Training von drei verschiedenen Korpora: SwissDial, SDS-200 und Korpus des Schweizer Parlaments (SPC) [33]. Um die semantische Distanz zwischen vorhergesagtem und tatsächlichem Wort zu messen, entwickelten sie eine eigene Loss-Funktion [33]. Die besten Resultate erzielten sie mit dem Modell Whisper Large [33]. Interessanterweise erzielten sie auf dem SDS-200 die besten Resultate, obwohl dieser sowohl hinsichtlich Samples als auch der Anzahl Sprecher:innen pro Region nicht ausgewogen ist [33].

Im gleichen Jahr erschien eine Forschungsarbeit von C. Paonessa et al., die die Herausforderungen bei einem System untersuchte, das gesprochene schweizerdeutsche Dialekte zu geschriebenem Hochdeutsch übersetzt [34]. Dabei wurde analysiert, wie es sich auswirkt, wenn einzelne Dialekte beim Training weggelassen werden (Leave-one-out) [34]. Verwendet wurde der STT4SG-350-Korpus [34]. Dabei wurde bei einigen Dialekten festgestellt, dass der Ausschluss eines Dialekts im Training dazu führt, dass derselbe im Testset weniger gut erkannt wird, als wenn der vollständige Datensatz verwendet wird [34]. Am schlechtesten schneidet das Wallis ab, wenn es nicht Teil des Trainingssets ist. Die Dialekte Zürich und Innerschweiz sind weniger davon betroffen und profitieren somit am meisten vom Training mit anderen Dialekten [34]. Die Experimente wurden mit drei verschiedenen Modellen durchgeführt [34]. Bei Wav2Vec-XLS-R erfuhren die Dialekte, die im Trainingsset waren, keine Leistungseinbussen, wenn ein anderer Dialekt weggelassen wurde [34]. Dies sei auf die Leistung von Wav2Vec-XLS-R zurückzuführen, das sich mit wenigen Daten an neue Sprachen anpassen kann [34].

2.3.2 Vorgängerarbeiten der ZHAW

An der Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) wurden in den letzten Jahren im Rahmen von Projekt- und Bachelorarbeiten Experimente zur Sprach- beziehungsweise Dialekterkennung durchgeführt.

Reiser und Fivian entwickelten im Rahmen ihrer Bachelorarbeit im Jahr 2021 einen Klassfikator basierend auf Wav2Vec 2.0-XLSR-53. Ihr Ziel war es, die Leistungsfähigkeit des Modells unter der Verwendung von Datensätzen mit geringen Ressourcen zu untersuchen [35]. Dazu führten sie verschiedene Experimente durch, die unter anderem die Identifikation von Akzenten und Variationen der Längen der Samples umfassten. Für die Experimente verwendeten sie die englischen und spanischen Korpora aus Mozillas Common Voice [35]. Reiser und Fivian konnten keine eindeutige Aussage darüber treffen, welche Länge die Samples für das Training aufweisen sollen, um eine Leistungsverbesserungen zu erzielen [35].

Im selben Jahr führten Stucki und Randjelovic im Rahmen ihrer Projektarbeit Experimente mit wav2vec2-XLS-R und schweizerdeutschen Audiodateien aus dem SDS-200-Korpus durch [36]. Unter anderem versuchten sie, die Dialekte nach Kantonen zu klassifizieren, was sich allerdings als nicht zufriedenstellend herausstellte [36]. Eine Beobachtung war, dass Kantone mit einer geringer Anzahl Samples dazu neigten, niedrige F1-Scores zu erzielen [36]. Zudem vermuteten sie, dass die Grenzen der Dialekte nicht nach Kantonen klassifiziert werden sollten [36]. Daraufhin teilten sie die Dialekte auf vier Regionen auf und erreichten einen gewichteten F1-Score von 50.23 % und einen Macro-F1-Score von 45.96 % [36]. Die Gruppierung der Dialekte in grössere Regionen war somit erfolgreich. Wie die Dialekte für die Klassifizierung am besten gruppiert werden könnten, blieb offen [36].

In ihrer Bachelorarbeit im darauffolgenden Jahr führten Stucki und Randjelovic weitere Experimente mit wav2vec2-XLS-R durch und entwickelten zwei Systeme, indem sie die Pretrained-Models mit ungelabelten Daten weiter trainierten [37]. Der höchste erzielte gewichtete F1-Score belief sich auf 0.49 [37].

Der Hauptfokus der Bachelorarbeit zur Erkennung von Schweizer Dialekten unserer direkten Vorgänger, Claudio Frei und Philippe Schneider, in diesem Jahr lag auf der Evaluierung von neu verfügbaren Mittel wie dem kürzlich veröffentlichten STT4SG-350-Korpus und dem Modell Whisper [38]. Für die Klassifizierung der Dialekte orientierten sich Frei und Schneider an den sieben Regionen aus dem STT4SG-350-Korpus [38]. Die Modelle wav2vec und Whisper trainierten und evaluierten sie mit verschiedenen Konfigurationen [38]. Kleine Whisper-Modelle zeigten vielversprechende Ergebnisse und eine Kombination des SDS-200- und STT4SG-350-Korpus erzielte eine Steigerung in der Erkennung der Dialekte [38]. Sie stellten interessanterweise fest, dass Modelle, die auf dem unbalancierten SDS-200-Korpus trainiert wurden, höhere F1 Scores erreichten im Vergleich zu Modellen, die auf dem balancierten STT4SG-350-Korpus trainiert wurden [38]. Das beste Modell erreichte einen Micro-F1-Score von 64.72 %, wobei ein gemischter Korpus sowie Augmentation-Methoden angewendet wurden [38].

2.4 Infrastruktur und Technologie

Zumal diese Arbeit auf der Codebasis der Vorgängerarbeit aufbaut, wurde die Programmiersprache Python und das auf PyTorch basierende Toolkit SpreechBrain für die Programmierung übernommen. SpeechBrain ist frei zugänglich und bietet eine Reihe an Funktionalitäten, um gesprochene Sprache zu verarbeiten und ein auf einem Pretrained-Model basierendes Modell zu erstellen [39]. Dies vereinfacht die Entwicklung und spart Zeit. Besonders hervorzuheben ist, dass YAML-Dateien genutzt werden können, um sämtliche Parameter der Modelle festzulegen. Dies schafft Ordnung und der Quellcode muss nicht verändert werden, wenn Parameter angepasst werden sollen.

Um die Modelle zu trainieren, musste eine Infrastruktur eingerichtet werden, mit der auf eine GPU zugegriffen werden kann. Bei fast 150 GB Audiodaten und unter der Verwendung von neuronalen Netzwerken ist es nicht möglich, diese in vernünftiger Zeit auf einer durchschnittlichen CPU zu trainieren. Dank dem Openstack-Cluster APU, das die ZHAW ihren Student:innen kostenlos anbietet, konnte eine virtuelle Laufzeitumgebung (VM) mit 650 GB Speicher und Zugriff auf eine 92-nVidia-Tesla-T4-Grafikkarte eingerichtet werden. Damit die nötigen Abhängigkeiten wie Programmiersprache und Libraries nicht manuell installiert werden mussten, wurde ein Docker-Image erstellt und die Verarbeitung der Daten sowie das Training in Docker-Containern ausgeführt. Dies ermöglichte, die meisten Modelle in wenigen Tagen zu trainieren. Im Anhang findet sich eine Auflistung aller Trainingsdurchläufe und der dafür benötigten Zeit.

Sollten für spätere Arbeiten mehr Experimente in kurzer Zeit durchgeführt werden, ist die erwähnte Grafikkarte jedoch zu langsam. Dann empfehlen sich kostenpflichtige Cloud-Anbieter:innen, über die leistungsstärkere Grafikkarten gemietet werden können.

3. Experimente

Die folgenden Experimente widmen sich hauptsächlich Faktoren der automatischen Dialekterkennung, die unabhängig von der Modellarchitektur sind. So wird untersucht, wie die Zusammenstellung der Trainingsdaten die Resultate beeinflusst und welche Schwierigkeiten die Dialekterkennung mit sich bringt. Zunächst wird das Modell der Vorgängerarbeit und dessen Ergebnisse reproduziert. Bei gleichbleibender Architektur und Modellkonfiguration werden dann vier weitere Experimente durchgeführt mit gleichbleibendem Modell und unterschiedlich zusammengestellten Trainingsdaten. Es werden die F1-Scores und die Confusion-Matrizen verglichen sowie die Klassifizierung der Samples bezüglich Sprecher:innen und geographischer Lage geprüft.

3.1 Experiment 1: Reproduktion der Ergebnisse der Vorgänger

Das erste Experiment beinhaltet eine Reproduktion der Ergebnisse der Vorgängerarbeit von Philipp Schneider und Claudio Frei, die in Abschnitt 2.3.2 erwähnt ist. Die Reproduktion beinhaltet einerseits, den vorhandenen Quellcode wieder lauffähig zu machen und Optimierungspotenzial am Code zu identifizieren. Andererseits sollen die Ergebnisse eines Trainingslaufes mit einem der bestehenden Modelle genauer analysiert werden, um daraus weitere Experimente ableiten zu können. In einem ersten Abschnitt wird ausgeführt, wie die finalen Modelle von den Vorgängern aufgebaut sind, mit welchen Daten sie trainiert wurden und welche Hyperparameter festgelegt wurden. Im zweiten Abschnitt wird die Vorgehensweise beim Reproduzieren der Ergebnisse erläutert, welche Änderungen vorgenommen werden mussten und wo Verbesserungspotenzial im Code besteht. Im letzten Abschnitt werden schliesslich die Ergebnisse ausgewertet, die das wieder lauffähige Modell ergeben hat.

3.1.1 Vorhandenes Modell

Frei und Schneider präsentieren am Ende ihrer Arbeit zwei Versionen ihres besten Modells. Eine basierend auf Wav2Vec2-XLS-R-300M und eine zweite auf Whisper. Beide Versionen bauen auf der gleichen Architektur auf, die in Abbildung 8 im Detail gezeigt wird. Zunächst werden die Audiodaten dem Pretrained-Model Whisper beziehungsweise wav2vec übergeben, das eine latente Repräsentation der Daten generiert. Letztere werden einem linearen Layer übergeben, der Projector genannt wird [38]. Dann wird statistisches Pooling[5] angewendet. Im Classifier wird schliesslich in einem weiteren linearen Layer auf die gleiche Anzahl Knoten wie Klassen reduziert und mit einem Log-Softmax die logarithmischen Wahrscheinlichkeiten für die Dialektregionen ausgegeben. Im Gegensatz zu normalen Wahrscheinlichkeiten ist dadurch der Betrag der Werte sowie der Abstand zwischen den Werten grösser. Folglich werden falsche Vorhersagen stärker bestraft, wie in mehreren Foren und Blogs diskutiert wird [41][43]. Zudem hat Log-Softmax zwei klare Vorteile: Die Trainingslaufzeit wird reduziert und die Berechnungen der Gradienten ist numerisch stabiler [41][42].

Das Modell wird in der Abbildung 8 beispielhaft mit drei Klassen dargestellt. Die Zwischenwerte entsprechen in der Dimension der Output-Dimension des Pretrained-Models. Die Dimension der Feature-Vektoren wird also erst im Classifier reduziert. Wichtig für das Training ist ausserdem, dass sowohl die Parameter des Pretrained-Models als auch der zusätzlichen Layers angepasst werden. Da die Architektur jedoch relativ einfach gehalten ist, hängen die Ergebnisse des Modells hauptsächlich vom Finetuning des Pretrained-Models auf die jeweilige Sprache ab.

[5] Beim statistischen Pooling wird der Durchschnitt des Feature-Vektors mit der Standardabweichung desselben konkateniert [40].

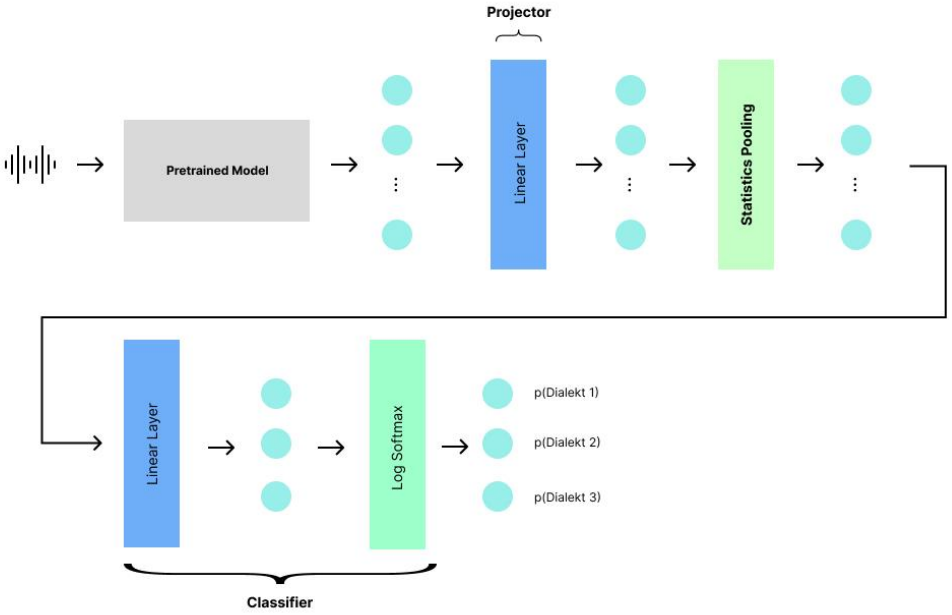


Abbildung 8: Architektur des Modells der Vorgängerarbeit gemäss Codeanalyse

Für das Training haben Frei und Schneider ein gemischtes Datenset aus dem SDS-200- und dem STT4SG-350-Korpus erstellt mit einer begrenzten Anzahl von Samples pro Dialektregion. Die besten Ergebnisse haben sie mit 10000 und 20000 Samples pro Dialektregion erzielt [38]. Weil sie vermuteten, dass ihre ersten Versuche statt Dialekte Sprecher:innen erkannt haben, haben sie mit Erfolg SpecAugment auf den Daten angewendet [38]. Diese Methode maskiert Teile der Frequenz und einzelne Zeitabschnitte der Audiodaten [44].

Bezüglich der Hyperparameter sind sie konservativ vorgegangen und haben sich bereits anfangs ihrer Arbeit mithilfe von Experimenten für Parameter entschieden und diese über die Experimente hinweg konstant gehalten [38]. So konnten sie sich auf andere Faktoren für das Training fokussieren. Für das Pretrained-Model und für die weiteren Layers wurde je eine Learning-Rate verwendet [38]. Bei der Batch-Size ist hervorzuheben, dass sie aufgrund des beschränkten Speichers Gradienten-Akkumulation implementiert haben, um eine grössere resultierende Batch-Size zu erreichen [38]. In Tabelle 2 werden die Werte der wichtigsten Hyperparameter aufgelistet.

Hyperparameter	Werte
Initiale Learning Rate Pretrained Model	0.00003
Finale Learning Rate Pretrained Model	0.00000003
Initiale Learning Rate Model	0.001
Finale Learning Rate Model	0.000001
Batch Size	Meist 8 mal Akkumulationsfaktor 4 = 32
Epochen	Meist 20

Tabelle 2: Verwendete Hyperparameter und ihre Werte

Der Loss wird mit der Negative-Log-Likelihood-Loss-Funktion berechnet und als Optimizer wird Adam mit dessen Standardparametern eingesetzt [38]. Dies gilt sowohl für das Pretrained-Model als auch für die zusätzlichen Modell-Teile. Der Negative-Log-Likelihood-Loss wurde ausgewählt, weil in der Architektur, wie bereits erwähnt, Log-Softmax verwendet wird [38]. Erst ganz am Ende ihrer Arbeit haben Frei und Schneider ein Hyperparameter-Tuning für ihr bestes Modell durchgeführt. Für alle Experimente dieser Arbeit werden jedoch bewusst die aufgeführten Hyperparameter genutzt, so dass mit einzelnen Beobachtungen der Vorgänger verglichen werden kann.

3.1.2 Vorgehen bei der Reproduktion

Bei der Übernahme des Quellcodes von Frei und Schneider wurde festgestellt, dass die Modelle so nicht lauffähig sind. Zu diesem Zeitpunkt war unklar, ob dies zum Beispiel an einer veralteten Code-Version lag oder ob sogar Komponenten fehlten. Deshalb wurden systematisch von den Input-Daten über die Datenaufbereitung bis hin zum Modell sämtliche Codeteile und Schritte überprüft. Parallel wurden Teile erneut implementiert mit punktuellen Optimierungen oder Alternativen, für die die Lauffähigkeit garantiert werden konnte.

In einem ersten Schritt werden die STT4SG-350-Daten im FLAC-Format neu kodiert, weil bei den Originaldaten Informationen fehlen, die es erlauben, die Audiodateien mit gängigen Mediaplayern abzuspielen [38]. Das entsprechende Skript hat wie gewünscht funktioniert und wurde nicht optimiert. Als zweiter Schritt werden aus den einzelnen Audiodateien im MP3- und FLAC-Format eine HDF5-Datei erstellt. Dies ermöglicht es dem Modell, effizient auf die Audioclips zuzugreifen und hat sich bereits in mehreren Arbeiten bewährt [38].[6] Auch da konnten keine Schwachstellen erkannt werden.

Im dritten Schritt werden die Daten normalisiert. Nicht verwendbare Samples werden entfernt und den verbleibenden Samples werden benötigte Metadaten hinzugefügt wie eine ID, zusätzliche Labels oder die Audioclip-Länge. In dem vorhandenen Skript wurde eine für beide schweizerdeutschen Korpora einzigartige ID mit einer Hash-Funktion hinzugefügt. Hash-Funktionen können in seltenen Fällen zu

[6] Im ReadMe von Frei und Schneider wird die Normalisierung als zweiter Schritt deklariert. Das entsprechende Skript muss allerdings auf die zu dem Zeitpunkt noch nicht erstellte HDF5-Datei zugreifen, um die Audio-Clip-Länge zu berechnen. Umgekehrt ist das Verpacken der Audiodateien unabhängig von der Normalisierung. Daher wird davon ausgegangen, dass es sich um ein Versehen im ReadMe handelt.

Kollisionen führen, so dass möglicherweise zwei Samples die gleiche ID erhalten. Zumal alle Fehlerquellen ausgeschlossen werden mussten – seien sie noch so unwahrscheinlich –, wurde in einem eigenen Skript eine andere Strategie entwickelt. Da die vorliegende STT4SG-350-Korpus-Version keine IDs enthält, wurden Client ID (ID der:des Sprecher:in) und Audioclip-Dateinamen zu einer ID konkateniert. Diese ID beinhaltet sowohl Buchstaben als auch Zahlen. Dies hat den Vorteil, dass sie sich von der ID des SDS-200-Korpus unterscheidet, bei dem die Samples von 0 weg durchgehend nummeriert sind. Demnach entfällt eine zusätzliche ID über beide Korpora hinweg.

In diesem Schritt wurde schliesslich auch das Problem entdeckt. Das Normalisierungs-Skript griff auf Dateien mit Korpus-Metadaten zu, die in der vorliegenden Korpus-Versionen nicht vorhanden waren. Auf Nachfrage konnten Frei und Schneider diese Dateien nachliefern, damit diese mit den vorliegenden Metadaten verglichen werden konnten. Daraus konnte geschlossen werden, dass es sich um verschiedene Korpora-Versionen handelt und lediglich das Normalisierungs-Skript auf die vorliegenden Versionen angepasst werden musste, um das Modell laufenlassen zu können. Damit garantiert werden konnte, dass das erklärte Problem das einzige ist und weiteres Optimierungspotenzial ausgeschöpft werden konnte, wurden auch die restlichen Schritte des Preprocessings überprüft.

Nach der Normalisierung kann optional die Anzahl der Samples reduziert werden, um überrepräsentierte Dialektregionen besser ausbalancieren zu können. Schneider und Frei beginnen mit einem leeren Datenset und füllen die einzelnen Regionen iterativ auf, indem sie abwechslungsweise von jede:r Sprecher:in ein Sample hinzufügen. So werden Samples von überrepräsentierten Sprecher:innen weggelassen. In einem eigenen Skript wurde ein umgekehrtes, erweitertes Verfahren entwickelt. Das bestehende Datenset wird iterativ reduziert, indem von den Sprecher:innen mit den meisten Samples Samples entfernt werden. Dies führt zwar zum gleichen Resultat, spart jedoch Zeit. Denn es müssen weniger Samples abgearbeitet werden und es muss nicht über die Sprecher:innen iteriert werden. Zusätzlich ist es beim neuen Skript möglich, dass der Datensatz auf die durchschnittliche Anzahl Samples pro Dialektregion reduziert wird, wenn nichts anderes angegeben wird.

Als letzter Schritt müssen Trainings-, Validierungs- und Testsplit erstellt werden. Weil es sich um einen für das Training kritischen Schritt handelt, wurde hier auch ein eigenes Skript erstellt. Die Herausforderung bei der Aufteilung der Datensätze in Splits besteht darin, dass einerseits keine Überlappung der Sprecher:innen in den Datensets auftreten und andererseits die ursprüngliche Verteilung der Daten beibehalten wird. Darüber hinaus sollen die Samples gemäss einem gewünschten Verhältnis aufgeteilt werden[7]. Gängige Methoden bieten keine Funktionen an, die alle oben genannten Kriterien berücksichtigen. Daher wurde sowohl bei Frei und Schneider als auch im eigenen Skript ein iteratives Split-Verfahren mithilfe einer gängigen Methode der Bibliothek scikit-learn implementiert [45]. Es werden mehrere Splits durchgeführt und derjenige Split ausgewählt, der dem gewünschten Verhältnis und der ursprünglichen Verteilung des Datensatzes am nächsten liegt. Das eigene Skript ist leicht vereinfacht, etwas anders aufgebaut und enthält am Schluss eine Testfunktion, die prüft, ob alle Regionen in allen Splits enthalten sind und keine Überschneidung der Sprecher:innen vorliegt. Sollte dies nicht der Fall sein, wird eine Warnung ausgegeben.

[7] Je nach Verteilung der Samples pro Sprecher:in und Anzahl Samples pro Region können diese Kriterien kaum oder nur teilweise erfüllt werden.

Beim Modell selbst wurden nur für die Lesbarkeit Code-Teile entfernt, die nicht zum finalen Experiment von Schneider und Frei gehörten, und nicht mehr benötigt wurden.

3.1.3 Set-up

Aus dem letzten Abschnitt folgt, dass nur die Anpassungen im Normalisierungs-Skript notwendig sind, um das Preprocessing und einen ersten Trainingsdurchgang durchführen zu können. Obwohl für fast jedes Skript eine punktuell optimierte Alternative erstellt wurde, wurden bis auf diese Anpassungen die originalen Skripte verwendet. Ziel war, damit die Resultate von Frei und Schneider möglichst genau reproduzieren zu können. Folglich wurden auch die Hyperparameter, wie in Abschnitt 3.1.1 beschrieben, übernommen. Als Labels werden wie in der Vorgängerarbeit die im STT4SG-350-Korpus verwendeten sieben Dialektregion genutzt, die bei der Normalisierung mithilfe eines Vergleichs der Postleitzahlen auf den SDS-200-Korpus gemappt werden, um die beiden Korpora mischen zu können.

Frei und Schneider haben die besten Ergebnisse erreicht, indem sie den SDS-200- und den STT4SG-350-Korpus kombiniert und die Anzahl Samples pro Dialektregion auf 10000 oder 20000 reduziert haben. Daher wurden auch für dieses Experiment die beiden Korpora gemischt, auf 10000 Samples pro Dialektregion reduziert und gesplittet. In Abbildung 9 ist die Verteilung der Samples und Sprecher:innen auf die Dialektregionen dargestellt. Wie erwartet, sind die Anzahl Samples pro Dialektregion ausgeglichen auf 10000. Die Anzahl Sprecher:innen variiert aber stark.

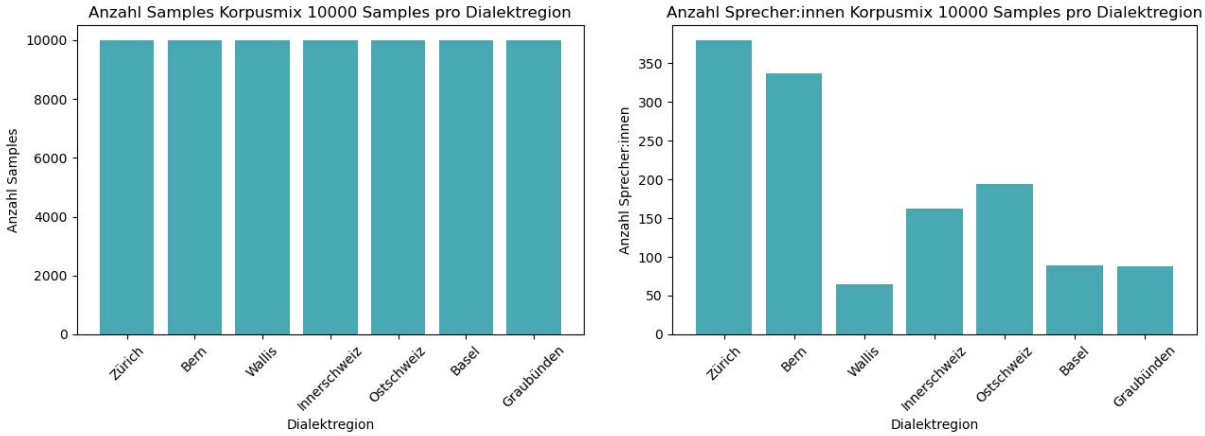


Abbildung 9: Statistiken zum Korpusmix mit 10000 Samples pro Dialektregion

Ein ähnliches Bild zeigt sich bei den drei Splits, wie auch Abbildung 10 zu sehen ist. Wobei die Anzahl Samples pro Region im Validierungs- und Testingset etwas unausgeglichener ist verglichen mit dem Trainingsset.

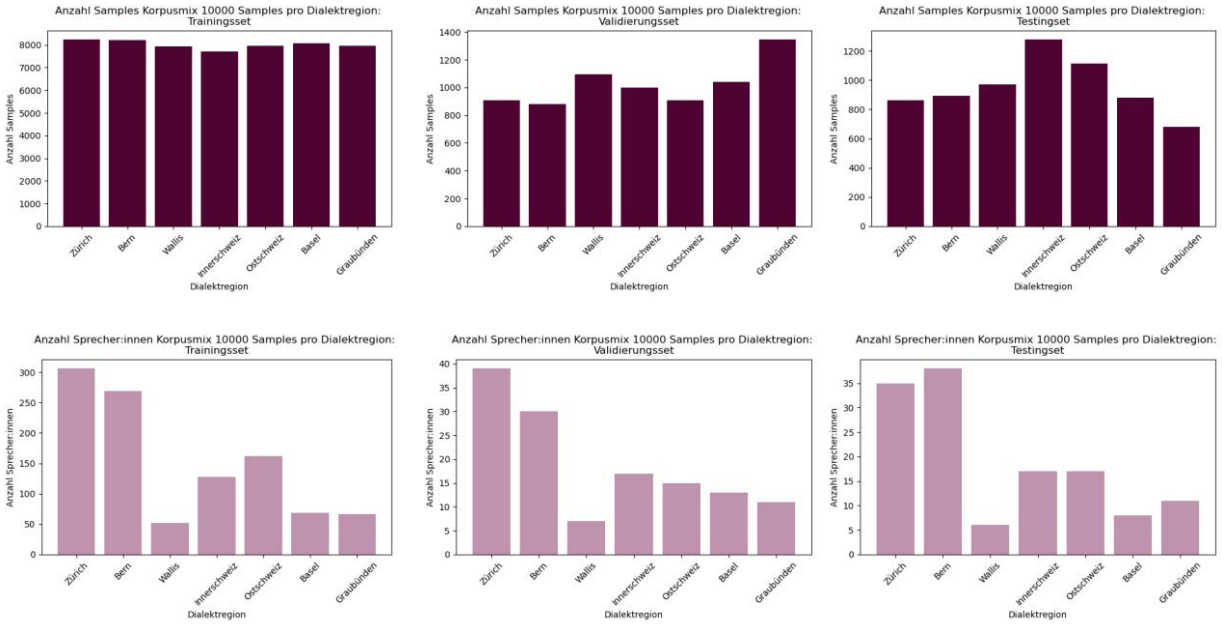


Abbildung 10: Statistiken zu den Splits für den Korpusmix mit 10000 Samples pro Dialektregion

Für dieses und alle folgenden Experimente wird nichts an der Modell-Architektur und den Hyperparameter geändert, damit die Experimente vergleichbar bleiben.

3.1.4 Analyse der Ergebnisse

Nach einem ersten Trainingslauf mit wav2vec, dessen Validierungsergebnisse bereits viel versprachen, wurde versucht, das Modell mit dem Testingset zu evaluieren. Dies lieferte jedoch sehr schlechte Scores und eine Confusion-Matrix, die darauf hindeutet, dass das Modell keine Dialekte unterscheiden kann. Schneider und Frei haben dafür ein von Trainings-Skript separates Testing-Skript erstellt, das Gefahr läuft, das gespeicherte Modell nicht richtig laden zu können. Einfachheitshalber wurde darum die Evaluation am Testingset dem Training im gleichen Skript nachgeschaltet. Für einen zweiten Trainingsdurchlauf wurde mit Whisper Base die kleinste Whisper-Version gewählt, die auch für alle kommenden Experimente genutzt wird. Obwohl Frei und Schneider etwas bessere Ergebnisse mit wav2vec erzielten, schnitt Whisper Base fast genauso gut ab, ist jedoch schneller und deshalb mit begrenzten Rechenressourcen praktischer [38]. Wie in Abschnitt 2.1.4 erwähnt, ist Whisper auch robuster gegen Hintergrundgeräusche. Dieses zweite Training lieferte schliesslich mit Frei und Schneider vergleichbare F1-Macro- und Micro-Scores, die sich um 60 % bewegen [38][8]. Tabelle 3 führt die Scores beider Sets auf. Für das Validierungsset sind die F1-Scores für das gesamte Set ca. 1.5 % höher. Bei beiden Sets schneidet die Innerschweiz schlecht ab. Basel und Graubünden schneiden nur im Testingset schlechter ab.

[8] In der Vorgängerarbeit ist nicht klar ersichtlich, ob Frei und Schneider jeweils die Scores für das Validierungsset oder das Testingset aufführen. Teilweise stehen keine genauen Angaben oder «valid». In dieser Arbeit wird sicherheitshalber mit beiden Sets evaluiert.

	VALIDIERUNGSSET		TESTINGSET	
CLASS	f1-score	support	f1-score	support
INNERSCHWEIZ	0.4670	1001	0.4429	1277
BERN	0.7200	882	0.5631	893
WALLIS	0.8988	1094	0.6983	969
OSTSCHWEIZ	0.6321	907	0.6793	1116
ZÜRICH	0.5968	908	0.5227	862
BASEL	0.6689	1042	0.4936	882
GRAUBÜNDEN	0.7411	1344	0.3439	681
ACCURACY	0.6807	-	0.5445	-
MACRO AVG	0.6750	7178	0.5348	

Tabelle 3: Ergebnisse des Validierungs- und Testingsets zum Experiment mit dem Korpusmix von 10000 Samples pro Dialektregion

Die Confusion-Matrizen in Abbildung 11 widerspiegeln letztere beiden Beobachtungen. Zudem wird häufig mit Zürich, Bern, Innerschweiz und Ostschweiz verwechselt, die durch die meisten Sprecher:innen im gesamten Datenset vertreten sind. Gerade bei Zürich mit den meisten Sprecher:innen ist dies am auffälligsten.

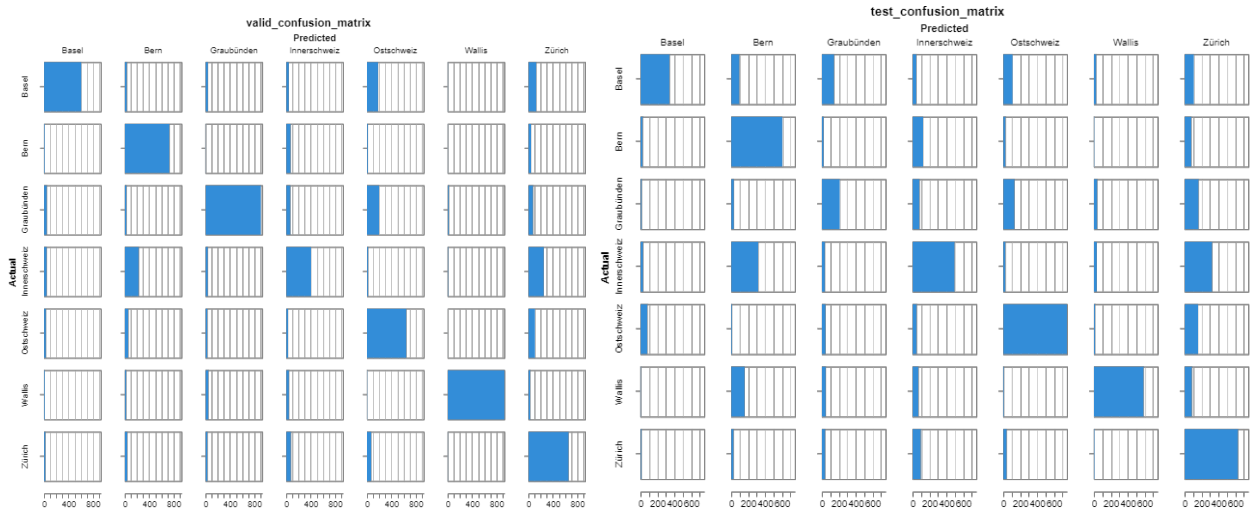


Abbildung 11: Confusion-Matrizen für die Resultate mit dem Korpusmix mit 10000 Samples pro Dialektregion

Mit Ausnahme von Graubünden sind die Confusion-Matrizen ähnlich verteilt. Die leichten Variationen in der Anzahl Samples lassen sich darauf zurückführen, dass die Dialektregionen jeweils im Validierungsset und dem Testingset nicht genau gleich viele Samples haben. Bei Graubünden muss genauer betrachtet werden, weshalb die Erkennung im Testingset so viel schlechter funktioniert. Wirft man einen Blick in die Statistiken des Testingsets in Abbildung 12, ist zu erkennen, dass zwei Sprecher:innen die Ergebnisse im

Testingset dominieren. Beide wurden schlecht vorhergesagt und sind gleichzeitig mit deutlich mehr Samples vertreten als die übrigen Sprecher:innen. Diese stehen also hinter dem Grossteil der falsch vorhergesagten Samples.

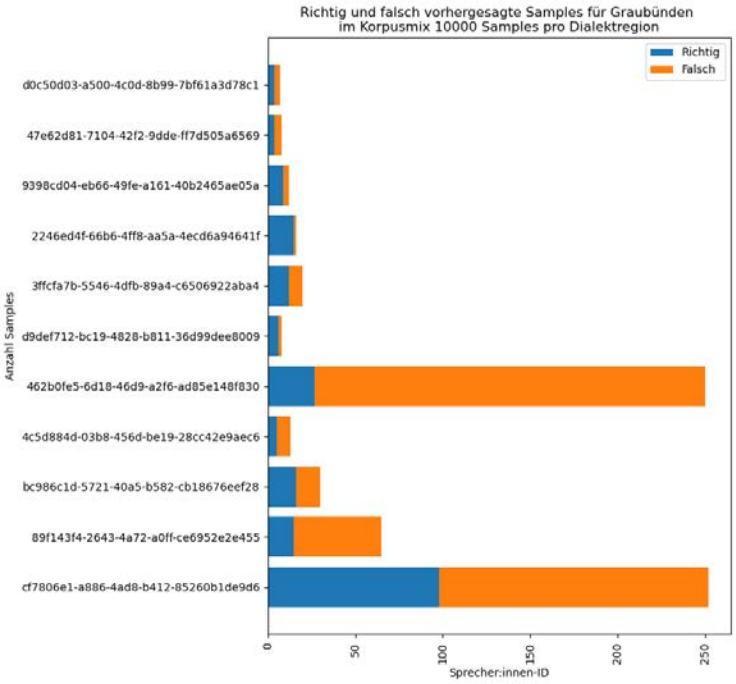


Abbildung 12: Richtig und falsch zugeordnete Samples nach Sprecher:innen in Graubünden

Bei einer Hörprobe bestätigt sich, dass der Dialekt der Sprecherin 462b0fe5-6d18-46d9-a2f6-ad85e148f830 aus dem SDS-200-Korpus nicht typisch nach Graubünden klingt und generell eher wenig ausgeprägte Dialektmerkmale aufweist, was ein Grund für die falsche Zuordnung sein könnte. Die im STT4SG-350-Korpus schlecht zugeordnete Sprecherin cf7806e1-a886-4ad8-b412-85260b1de9d6 klingt jedoch klar nach Graubünden. Es wurde jedoch bei einigen Proben ein Rauschen festgestellt, das einen negativen Einfluss auf die Vorhersage haben könnte. Die Anomalie bei Graubünden hat also vermutlich nichts mit dem Modell selbst zu tun, sondern liegt an der Audioqualität und den Sprecherinnen.

Da das Modell insgesamt doch bereits gute Resultate liefert, lohnt es sich, dies nochmals aus einer anderen Perspektive zu betrachten; nämlich aus der der Sprecher:innen. Zum einen war bereits bei Schneider und Frei ein Thema, dass das Modell statt Dialekten eigentlich Sprecher:innen erkennen könnte und schlechtere Ergebnisse liefert, weil es die Sprechermerkmale statt der Dialektmerkmale gelernt hat und diese bei neuen Sprecher:innen nicht mehr wiedererkennen kann [38]. Die Scores ihres finalen Modells lassen jedoch vermuten, dass dieses Problem deutlich reduziert wurde, die Dialekte schon gut vorhergesagt werden und die SpecAugmentation gut funktioniert. Zum anderen könnte spannend sein, wie klar das Modell ein:e Sprecher:in einem bestimmten Dialekt oder einer kleinen Auswahl von Dialekten zuordnet. So ist davon auszugehen, dass einige Sprecher:innen durch mehr als eine der Dialektregionen beeinflusst werden könnten.

Abbildung 13 stellt drei Heatmaps für die Sprecher:innen von Zürich, Bern und der Ostschweiz ab. Zürich hat im gesamten Datenset, wie erwähnt, knapp vor Bern am meisten Sprecher:innen. Beide Dialektregionen zeigen ein ähnliches Bild. So werden die Samples der meisten Sprecher:innen überwiegend der korrekten Dialektregion zugeordnet. Auch die Ostschweiz wird mit etwas weniger Sprecher:innen grösstenteils korrekt identifiziert. Die Anzahl Sprecher:innen könnte also einen positiven Einfluss auf die Zuordnung haben. Auch schwach erkennbar ist, dass die Regionen mit einigen anderen Regionen eher verwechselt werden.

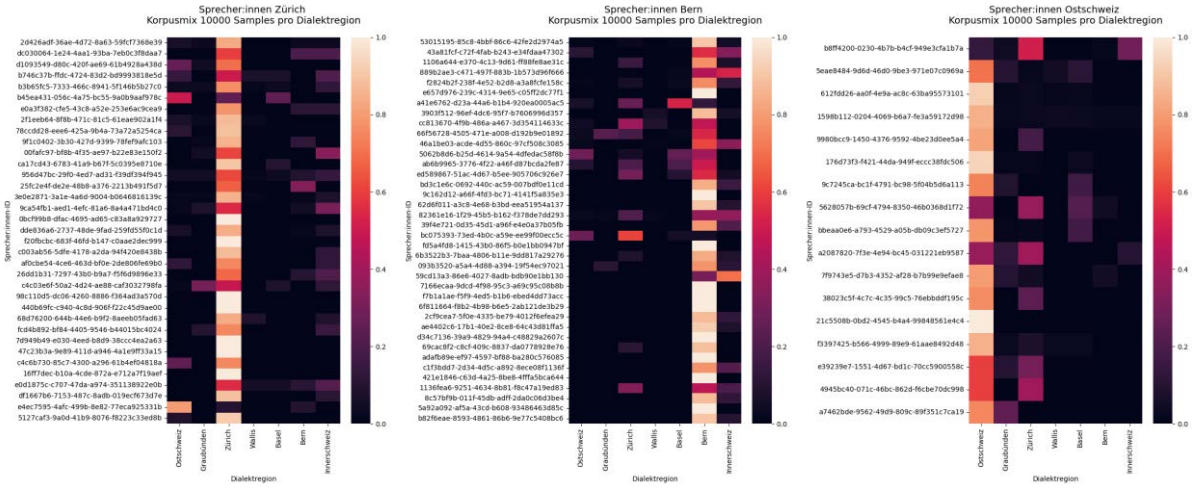


Abbildung 13: Heatmaps für die Zuordnung der Samples nach Sprecher:innen für Zürich, Bern und die Ostschweiz

Dieses Phänomen zeigt sich bei der Innerschweiz am stärksten. Einige Sprecher:innen werden klar richtig zugeordnet, andere werden bis zu drei Dialektregionen zu ähnlichen Anteilen zugeordnet. Zudem zeigt sich auch hier, was bereits in den Confusion-Matrizen in Abbildung 14 erkennbar ist, dass das Modell die Innerschweiz mit Zürich und Bern verwechselt.

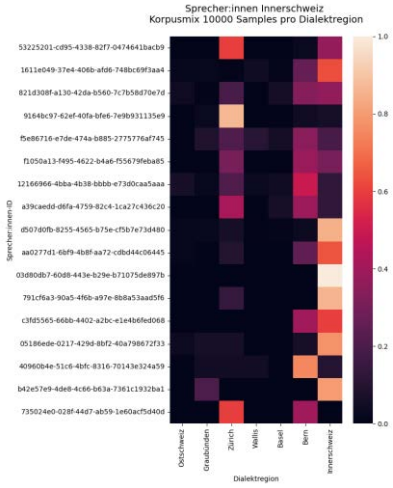


Abbildung 14: Heatmap für die Zuordnung der Samples nach Sprecher:innen für die Innerschweiz

Abbildung 15 stellt die Heatmaps der Dialektregionen mit den wenigsten Sprecher:innen im Trainingsset dar. Die Heatmap von Basel zeigt das unklarste Bild. Der Dialekt von Balser Sprecher:innen wird bis auf das Wallis mit allen Dialekten verwechselt und kann bei den meisten Sprecher:innen nicht klar einem Dialekt zugeordnet werden. Die Sprecher:innen von Graubünden werden besser und grösstenteils entsprechend Label zugeordnet. Fünf Sprecher:innen werden schlechter zugeordnet; darunter auch die beiden bereits besprochenen. Das Wallis mit den wenigsten Sprecher:innen im Trainingsset wirkt zwar auf den ersten Blick auch nicht positiv, doch fünf der sechs Sprecher:innen werden doch überwiegend dem Wallis zugeordnet und nur mit einer oder zwei Dialektregionen verwechselt. Was als gutes Ergebnis zu interpretieren ist.

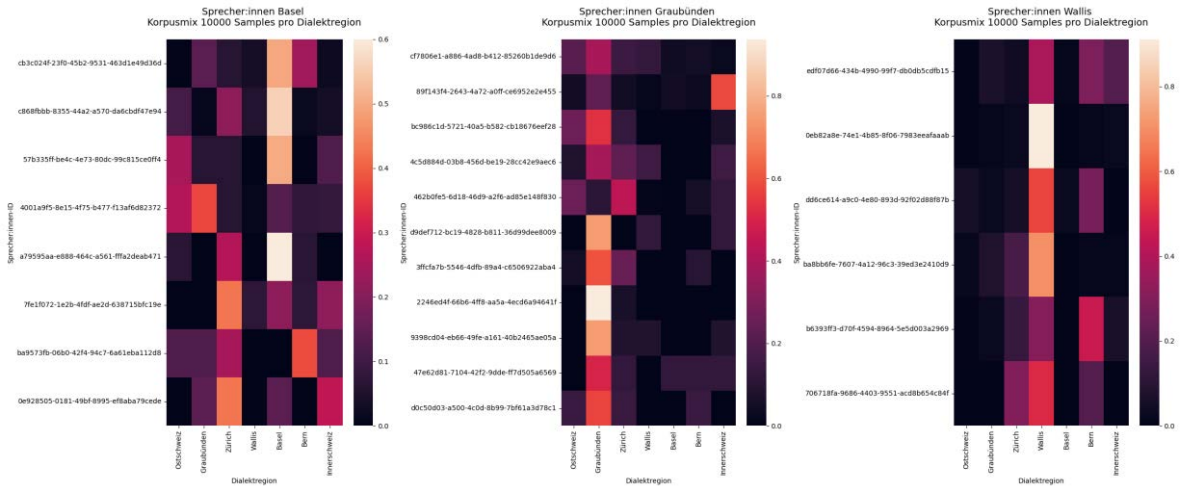


Abbildung 15: Heatmaps für die Zuordnung der Samples nach Sprecher:innen für Basel, Graubünden und das Wallis

Zusammengefasst liefert das Modell also für Basel die unklarsten Resultate und verwechselt die Innerschweiz am häufigsten. Über alle Heatmaps hinweg ist ein weiteres Phänomen zu beobachten. Vereinzelte Sprecher:innen werden gar nicht der korrekten Region zugeordnet, sondern einer oder zwei anderen. Wie bereits bei Graubünden gesehen, steht dies eher mit den Sprecher:innen direkt im Zusammenhang. Gerade wenn die Dialektregionen ansonsten korrekt vorhergesagt werden. Es könnte auf schlechte Audioqualität hindeuten, auf einen Fehler in den Metadaten wie beispielsweise bei der Postleitzahl[9] oder darauf, dass das Label nicht mit dem tatsächlichen Dialekt der Sprecher:innen übereinstimmt. Letzteres könnte daran liegen, dass die Sprecher:innen einen anderen Dialekt sprechen, als angenommen. Um mehr Klarheit zu schaffen, wurde eine Stichprobe der klar falsch vorhergesagten Sprecher:innen einzeln untersucht. In Tabelle 4 werden diese aufgelistet. Bei sechs von neun Sprecher:innen wurde beobachtet, dass die geographischen Orte hinter den Postleitzahlen verglichen zu den anderen Samples und Sprecher:innen eher am Rand der jeweiligen Region liegen oder sogar in einem Gebiet, in dem sich Samples zweier oder dreier Regionen überlappen (grün eingefärbt). Bei den restlichen drei

[9] Beim SDS-200-Korpus wurde, wie in Abschnitt 3.1.2 erklärt, die Dialektregion mithilfe der Postleitzahlen zugeordnet. Falls die Sprecher:innen versehentlich statt der Postleitzahl ihres Herkunftsorts, die ihres aktuellen Wohnorts angegeben haben, könnte dies zu einem falschen Label führen.

Sprecher:innen liegt der Ort der entsprechenden Postleitzahl geographisch nicht in einem überlappenden Gebiet (rot eingefärbt).

Sprecher ID	PLZ	Korpus	Dialektregion	Meist identifizierte Dialektregion(en)
b8ff4200-0230-4b7b-b4cf-949e3cfa1b7a	9442	STT4SG-350	Ostschweiz	Zürich
e4ec7595-4afc-499b-8e82-77eca925331b	8450	SDS-200	Zürich	Ostschweiz
b45ea431-056c-4a75-bc55-9a0b9aaf978c	8610	SDS-200	Zürich	Ostschweiz
b6393ff3-d70f-4594-8964-5e5d003a2969	3998	SDS-200	Wallis	Bern
ba9573fb-06b0-42f4-94c7-6a61eba112d8	4460	SDS-200	Basel	Bern
0e928505-0181-49bf-8995-ef8aba79cede	4402	SDS-200	Basel	Zürich
a41e6762-d23a-44a6-b1b4-920ea0005ac5	1700	SDS-200	Bern	Basel
bc075393-73ed-4b0c-a59e-ee99f00ecc5c	4143	SDS-200	Bern	Zürich
59cd13a3-86e6-4027-8adb-bdb90e1bb130	4616	SDS-200	Bern	Innerschweiz

Tabelle 4: Stichprobe der falsch vorhergesagten Sprecher:innen zum Experiment mit dem Korpusmix

Neun Sprecher:innen sind jedoch nicht repräsentativ. Darum wurde die geographische Lage und Korrektheit der Vorhersagen nochmals für alle Sprecher:innen unter die Lupe genommen. Als Vergleichsbasis dient Abbildung 16, die die Sprecher:innen der beiden ganzen Korpora gemäss der angegebenen Postleitzahl auf einer Landkarte darstellt. Bei transparenteren Punkten sind weniger Sprecher:innen beziehungsweise Samples der jeweiligen Dialektregion verortet, während Häufungen mit kaum oder keiner Transparenz hervortreten.

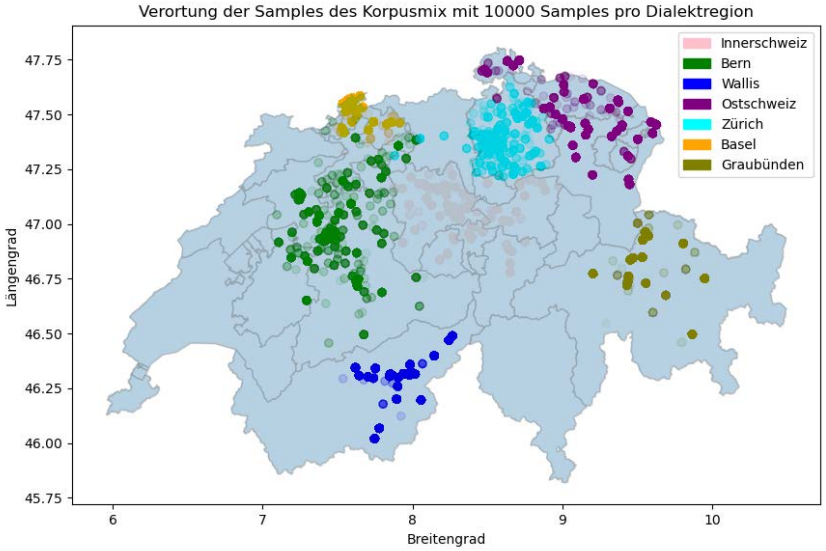


Abbildung 16: Geographische Verortung der Samples des Korpusmix mit 10000 Samples pro Dialektregion

Bern, das gut vorhergesagt wird und durch viele Sprecher:innen vertreten ist, wird als Beispiel in Abbildung 17 auf der Landkarte verortet. Links wird die Farbabstufung entsprechend der Heatmaps nochmals gezeigt (je heller, desto besser vorhergesagt) und rechts wird gezeigt, welcher Region die Sprecher:innen am häufigsten zugeordnet wurden. Im Zentrum von Bern werden die Sprecher:innen am besten zugeordnet. Gegen aussen nimmt die Anzahl Samples ab, die der korrekten Region zugeteilt wurden. Bei den mehrheitlich falsch zugeordneten Sprecher:innen, die ganz aussen liegen, macht die Verwechslung mit der Innerschweiz geographisch am meisten Sinn, die mit Zürich ist nicht komplett falsch, die mit Basel erscheint jedoch nicht sinnvoll.

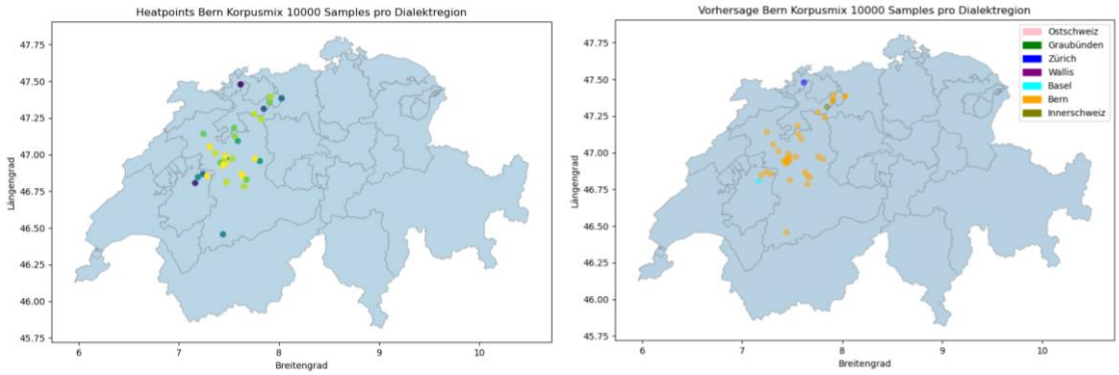


Abbildung 17: Heatpoints und vorhergesagte Labels für Bern

Die hier nicht abgebildeten, gut erkannten Dialektregionen verhalten sich ähnlich (siehe Anhang). Dies könnte bedeuten, dass das Modell Sprecher:innen nicht gut einordnet, wenn diese potenziell Merkmale von zwei oder mehr angrenzenden Kantonen mitbringen. Weil sich die Schweizer Bevölkerung stark durchmischt, ist sogar denkbar, dass Sprecher:innen einen Mix aus Dialekten sprechen, die nicht aneinander angrenzen. Ob das Modell bereits genug genau ist, dass es solche Grenzfälle richtig identifiziert und die Samples den Dialektregionen zuordnet, die tatsächlich zu einem solchen Mix beitragen, ist in Frage zu stellen. Dies kann auch nicht überprüft werden, da nur ein Label pro Sample verfügbar ist.

3.1.5 Fazit

Die Reproduktion der Resultate von Frei und Schneider ist gelungen und es konnten ähnliche Resultate erzielt werden. Die Analyse der Ergebnisse zeigt, dass die Innerschweiz noch nicht gut vorhergesagt wird. Eine Erklärung könnte sein, dass sich die Merkmale der Innerschweiz eher mit anderen Dialektregionen überschneidet, weil es an die meisten anderen Regionen grenzt. Basel und Graubünden liefern nur mit dem Testingset schlechte Ergebnisse, wobei sich dies bei Graubünden anhand von zwei dominanten Sprecher:innen erklären liess. Dialektregionen mit mehr Sprecher:innen schneiden sowohl über alle Samples hinweg als auch für einzelne Sprecher:innen besser ab, führen aber zu häufigeren Verwechslungen bei anderen Regionen. Im Weiteren soll diese Dysbalance genauer untersucht werden, um auszuloten, wie das Trainingsset am besten zusammengestellt werden sollte.

Zudem konnte gezeigt werden, dass Samples von Sprecher:innen nahe beim geographischen Zentrum der Dialektregionen tendenziell besser zugeordnet werden können als Samples von Sprecher:innen, die eher am Rand der Regionen beheimatet sind. Dabei wurden die Samples jedoch nur teilweise benachbarten

Dialektregionen zugeordnet. Wie erklärt, kann so nicht entschieden werden, ob das Modell tatsächlich Merkmale eines nicht dem Label entsprechenden Dialektes korrekt identifiziert oder schlicht falsch liegt. Sobald das Modell in folgenden Arbeiten verbessert werden kann, könnte man zu dieser Frage zurückkehren und prüfen, ob sich dies immer noch wie beschrieben verhält. Dies könnte bedeuten, dass sich bei einem solchen Modell die obere Grenze potenzieller Scores nach unten verschieben würde, je ambiger die Samples sind.

3.2 Experiment 2: Balancierter STT4SG-350-Korpus

In diesem Experiment wird der SST4SG-350-Korpus verwendet, der hinsichtlich der Anzahl Sprecher:innen zwar ausgewogen ist, im Gegensatz zum im vorherigen Experiment verwendeten Korpusmix aber insgesamt weniger Sprecher:innen beinhaltet. Wie der Korpusmix ist auch der SST4SG-350-Korpus bezüglich Anzahl Samples pro Region ausgeglichen. Dies ermöglicht es, zu untersuchen, welchen Effekt eine höheren Anzahl Samples und gleichzeitig kleinere Anzahl Sprecher:innen hat.

3.2.1 Set-up

Bei diesem Experiment wurden keine eigenen Splits erstellt, sondern die vom SST4SG-350-Korpus zur Verfügung gestellten Splits übernommen [21]. Die Splits wurden gemäss dem in Abschnitt 3.1.2 beschriebenen Preprocessing-Verfahren bereinigt. Das Validierungsset verfügt über etwas weniger Samples als das Testingset, wie in Abbildung 18 ersichtlich. Was die Anzahl der Sprecher:innen angeht, enthält das Validierungsset ungefähr einen Drittel weniger Sprecher:innen als das Testingset.

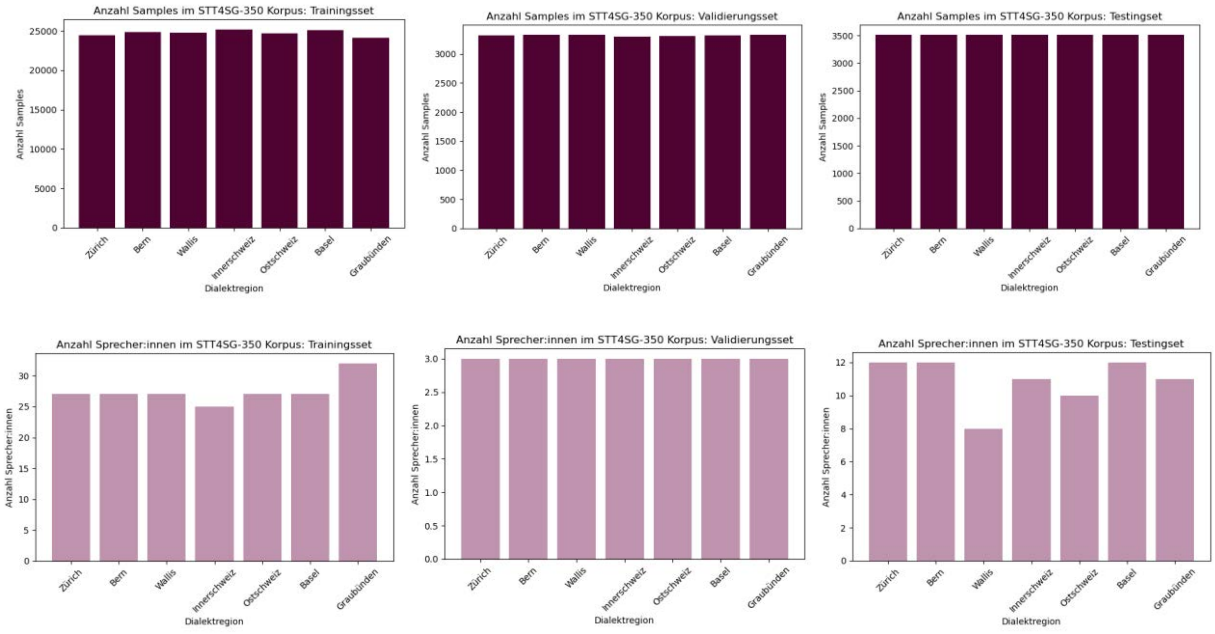


Abbildung 18: Statistiken zu den vom STT4SG-350-Korpus zur Verfügung gestellten Splits

3.2.2 Analyse der Ergebnisse

Im Vergleich zum vorherigen Experiment mit dem Korpusmix fallen die Ergebnisse im Testingset schlechter aus, wie die Scores in Tabelle 5 zeigen. Lediglich 75 % des Macro-F1-Scores konnten im Vergleich zum ersten Experiment beibehalten werden. Die Erkennung der Dialekte Zürich und Bern ist wesentlich schwächer, während Graubünden besser abschneidet. Letzteres hängt wohl damit zusammen, dass die Sprecher:innen im Testingset des STT4SG-350-Korpus nicht enthalten sind, die die Resultate im Korpusmix zuvor negativ beeinflusst haben.

Der Korpusmix enthielt mehr Sprecher:innen für die geografisch angrenzenden Regionen Zürich und Bern, mit denen die Innerschweiz häufig verwechselt wurde. Durch die geringere Anzahl der Sprecher:innen aus Zürich und Bern im SST4SG-350-Korpus sinkt die Wahrscheinlichkeit für diese Verwechslung. Zwar wird die Innerschweiz im Vergleich zum Korpusmix besser erkannt, dennoch werden ihr viele Dialekte fälschlicherweise zugeordnet, insbesondere aus den angrenzenden Regionen Zürich und Bern. Es scheint, dass diese Regionen schwer voneinander zu unterscheiden sind. Die Ähnlichkeit zwischen den Dialekten Innerschweiz und Zürich wurde bereits in der Forschungsarbeit von Paonessa et al. beobachtet, die in Abschnitt 2.3.1 erwähnt wird.

Vergleiche der Resultate innerhalb dieses Experiment zeigen, dass der Macro-F1-Score im Validierungsset insgesamt niedriger ausfällt als im Testingset (vgl. Tabelle 5). Dies könnte daran liegen, dass im Validierungsset weniger Sprecher:innen enthalten sind. Diese Beobachtung deckt sich mit derjenigen aus dem vorhergehenden Experiment und stärkt die Annahme, dass die Anzahl der Sprecher:innen einen Einfluss auf die Ergebnisse hat.

Die Innerschweiz weist den grössten Unterschied zwischen Validierungsset und Testingset auf, wobei Samples im Testingset offenbar schwieriger zu erkennen sind. In beiden Sets werden allerdings viele Samples von Bern und Zürich der Innerschweiz zugeordnet.

	VALIDIERUNGSSET		TESTINGSET	
CLASS	**f1-score**	**support**	**f1-score**	**support**
ZÜRICH	0.3432	3312	0.2334	3515
INNERSCHWEIZ	0.0886	3296	0.3950	3515
GRAUBÜNDEN	0.3759	3328	0.4723	3515
BASEL	0.2219	3322	0.4011	3515
OSTSCHWEIZ	0.2127	3302	0.5159	3515
WALLIS	0.3621	3328	0.5687	3515
BERN	0.3229	3329	0.2779	3515
ACCURACY	0.2726	-	0.4093	-
MACRO AVG	0.2753	23217	0.4092	24605

Tabelle 5: Ergebnisse des Validierungs- und Testingsets für das Experiment mit dem balancierten STT4SG-350-Korpus

Dennoch ist an der Diagonalen der Confusionmatrix in Abbildung 19 für das Testingset deutlich ersichtlich, dass das Modell Dialekte erkennt.

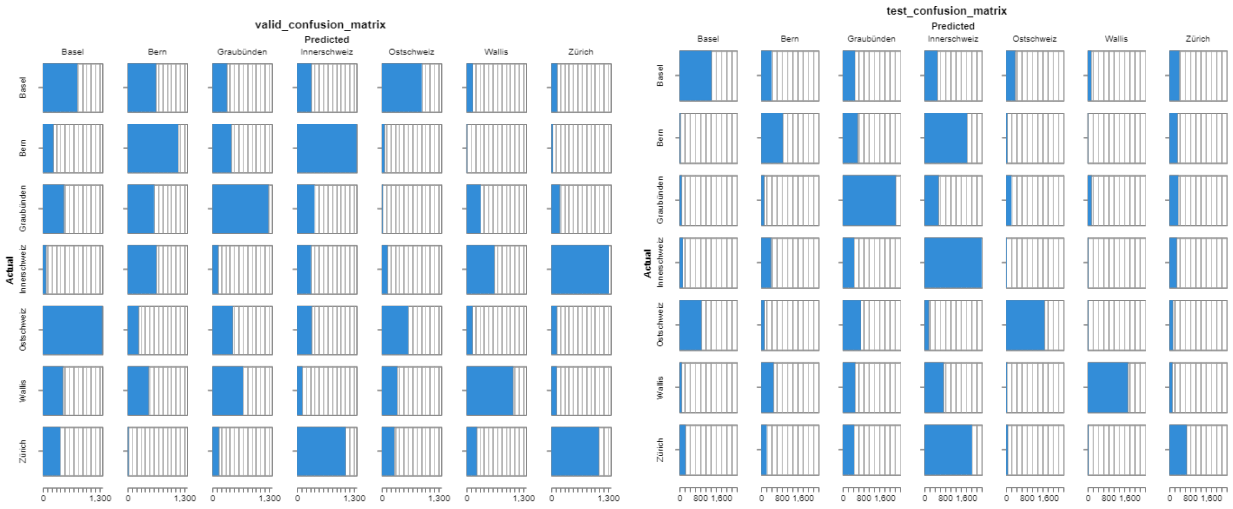


Abbildung 19: Confusion-Matrizen zu den Vorhersagen am ganzen STT4SG-350-Korpus

Da im vorhergehenden Experiment bei der Betrachtung der Sprecher:innen bereits erfolgreich eine Erklärung für die schlechten Resultate bei Graubünden gefunden werden konnte, werden Sprecher:innen, deren Samples mehrheitlich und fälschlicherweise der Innerschweiz zugeordnet wurden, in der Abbildung 20 durch die hellen Rechtecke dargestellt, näher betrachtet.

Die Sprecher:in 887b50f8-215b-4a1d-8f32-13516da6506f in ihren Sechzigern spricht, wie gelabelt, Berndeutsch. Der junge Sprecher 8b48025a-2cff-4346-9dab-b45cfb683e22 spricht Zürichdeutsch, ebenfalls entsprechend Label. Der Sprecher ba118975-5963-4495-927a-a78d19dd98c1 in seinen vierziger Jahren spricht ein undeutliches Berndeutsch. Letztere Aufnahme ist zudem mit einem Rauschen unterlegt. Diese einzelnen Sprecher:innen weisen folglich keine gemeinsamen Auffälligkeiten in den Aufnahmen oder ihren Metadaten auf, die die Fehlzuordnungen erklären könnte. Diese Stichprobe zeigt, dass die Gründe für die Fehlzuordnungen, zumindest in diesen Fällen, nicht bei den Sprecher:innen liegt.

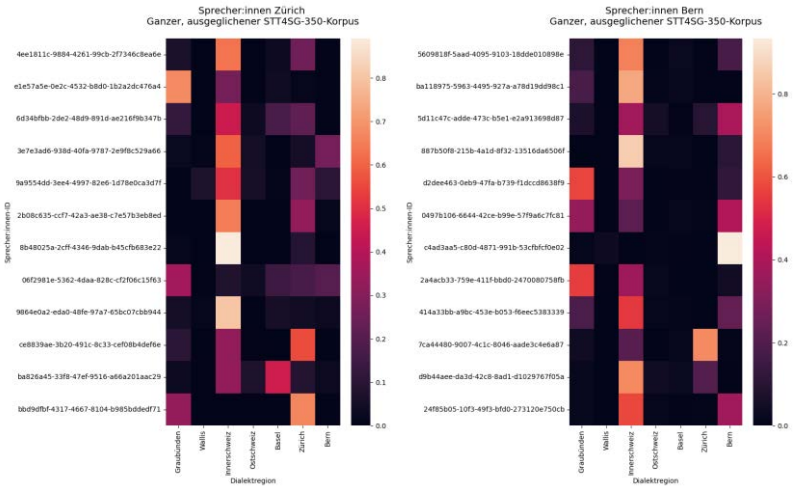


Abbildung 20: Heatmaps für die Zuordnung der Samples nach Sprecher:innen für Zürich und Bern

Die Heatmaps in Abbildung 21decken sich mehr mit denen aus dem ersten Experiment. Basel wird zwar tendenziell etwas besser vorhergesagt als beim Korpusmix. Insgesamt werden die Sprecher:innen jedoch nahezu jeder Region zugeordnet. Die Samples der Sprecher:innen in der Ostschweiz und dem Wallis werden klar besser identifiziert.

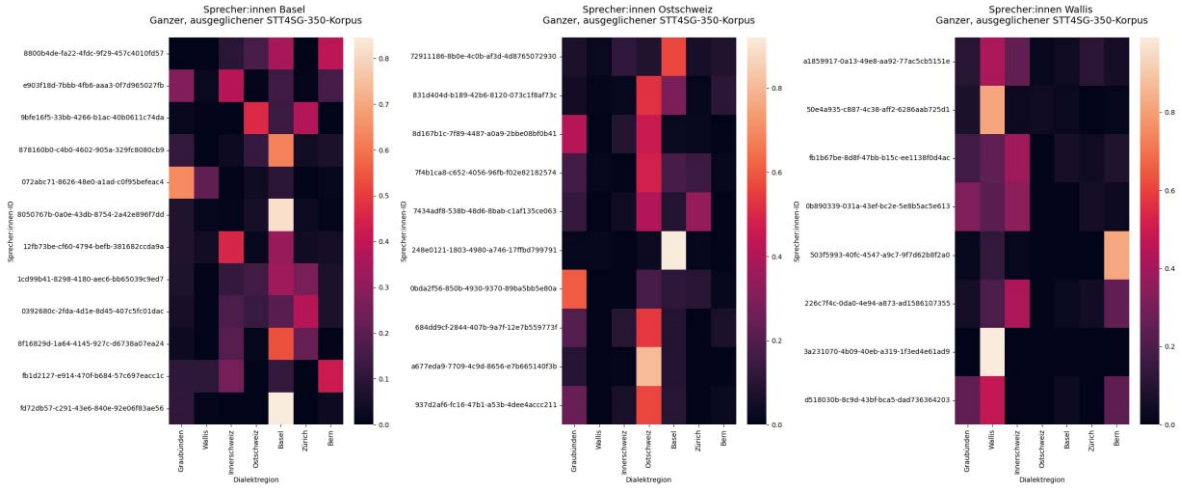


Abbildung 21: Heatmaps für die Zuordnung der Samples nach Sprecher:innen für Basel, die Ostschweiz und das Wallis

3.2.3 Fazit

Es scheint, dass die Innerschweiz, Zürich und Bern nur schlecht voneinander trennbar sind, da die Regionen sowohl im Modell des ersten Experiments als auch in diesem Experiment zu Verwechslungen geführt haben, wenn auch in entgegengesetzte Richtungen. Der Korpusmix schnitt trotz weniger Trainingssamples (54525 Samples) besser ab als der STT4SG-350-Korpus mit 173194 Samples. Umgekehrt hat der STT4SG-350-Korpus mit 192 Sprecher:innen weniger Sprecher:innen als der Korpus mit 1052

Sprecher:innen, insbesondere was die Anzahl Sprecher:innen in den Regionen Zürich und Bern betrifft. Daraus bestätigt sich die Annahme, dass die Anzahl der Sprecher:innen einen entscheidenden Einfluss auf die Leistung des Modells hat.

3.3 Experiment 3: Englischer Korpus

Obwohl Whisper multilingual trainiert wurde, ist doch nur ein Drittel der Daten nicht Englisch, wie in Abschnitt 2.1.4 erwähnt. Und diesen Drittel teilen sich 96 andere Sprachen (vgl. Abschnitt 2.1.4). Englisch ist also klar am stärksten in den Trainingsdaten repräsentiert. Dies legt die Vermutung nahe, dass Whisper am besten mit englischen Daten funktioniert. Um herauszufinden, ob die schweizerdeutschen Korpora deswegen einen Nachteil haben und ob die Modelle noch ein grösseres Potenzial mit englischen Daten haben, wurde für das folgende Experiment ein englischer Korpus hinzugezogen (vgl. Abschnitt 2.2.3). Ausserdem wird der englische Korpus auch für weitere Untersuchungen interessant sein bezüglich den Faktoren, die ausschlaggebend für die Dialekterkennung sind.

3.3.1 Set-up

Im Vergleich zu den schweizerdeutschen Korpora mit insgesamt über 300000 Samplen sind im englischen Korpus mit 17877 deutlich weniger Samples vorhanden. Mit 32 Sprecher:innen im Vergleich zu 316 Sprecher:innen vom STT4SG-350-Korpus oder 3997 Sprecher:innen im SDS-200-Korpus, sind es auch viel weniger Sprecher:innen. Betrachtet man Abbildung 22, ist zu erkennen, dass die Anzahl Samples pro Dialektregion nicht ausgeglichen ist. Eine ähnliche Verteilung gilt für die Anzahl Sprecher:innen pro Dialektregion (siehe Abbildung 22).

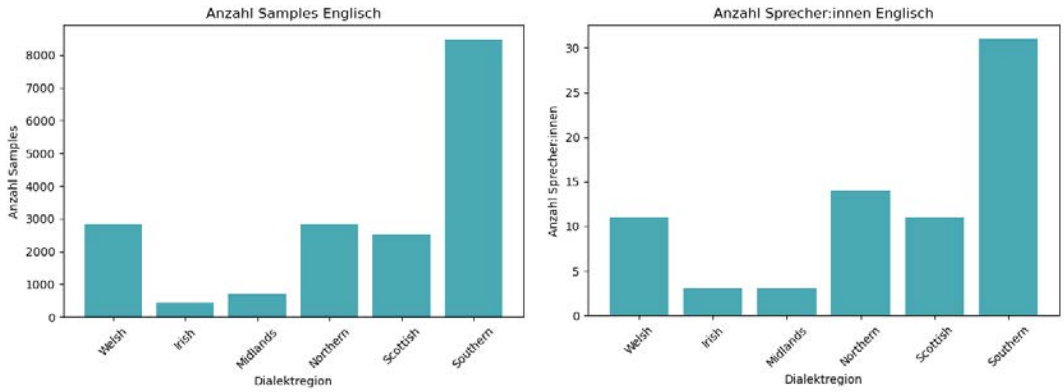


Abbildung 22: Statistiken zum englischen Korpus

Umgekehrt ist die Anzahl Samples pro Sprecher:in mit von 150 bis 300 Samples über alle Regionen hinweg ausgeglichen. Somit sinkt das Risiko, dass Merkmale eine:r Sprecher:in überrepräsentiert sind.

Da es insgesamt zu wenig Samples sind, wurde im Preprocessing zunächst darauf verzichtet, das Ungleichgewicht zwischen den Regionen auszugleichen. Es wäre jedoch denkbar, in einem weiteren

Experiment beim Southern-Dialekt Samples zu entfernen, der durch mehr als doppelt so viele Samples vertreten ist als Northern mit den zweitmeisten Samples.

Für das Erstellen der Trainings, Validierungs- und Testingsplits mussten wegen der geringen Sprecher:innenanzahl speziell eine neues Verfahren entwickelt werden. Denn ist die Anzahl der Sprecher:innen in einer Region sehr klein beispielsweise nur 3 und es soll ein Verhältnis von zum Beispiel 80:10:10 für die Splits erreicht werden, würden 0.8 * 3, also mindestens zwei Sprecher:inn, dem Trainingsset zugeteilt. Dies hätte zur Folge, dass in den übrigen Sets keine Sprecher:innen dieser Region und folglich auch keine Samples zugeteilt werden. Daher werden die Sprecher:innen einer Region mit Kombinatorik auf die Sets verteilt und die beste Kombination, die das gewünschte Verhältnis am besten erfüllt, ausgewählt. Da dieses kombinatorische Verfahren aufwendig ist, wird ab einer ausreichend grossen Anzahl Sprecher:innen in einer Region wieder auf die im Experiment 1: Reproduktion der Ergebnisse der Vorgänger erwähnte iterative Methode zurückgegriffen.

Entstanden sind die Sprecher:innen- und Sampleverteilungen in Abbildung 23. Die Verteilung der Sprecher:innen und Samples auf die Dialektregionen bleibt anteilmässig verglichen mit dem gesamten Korpus bestehen.

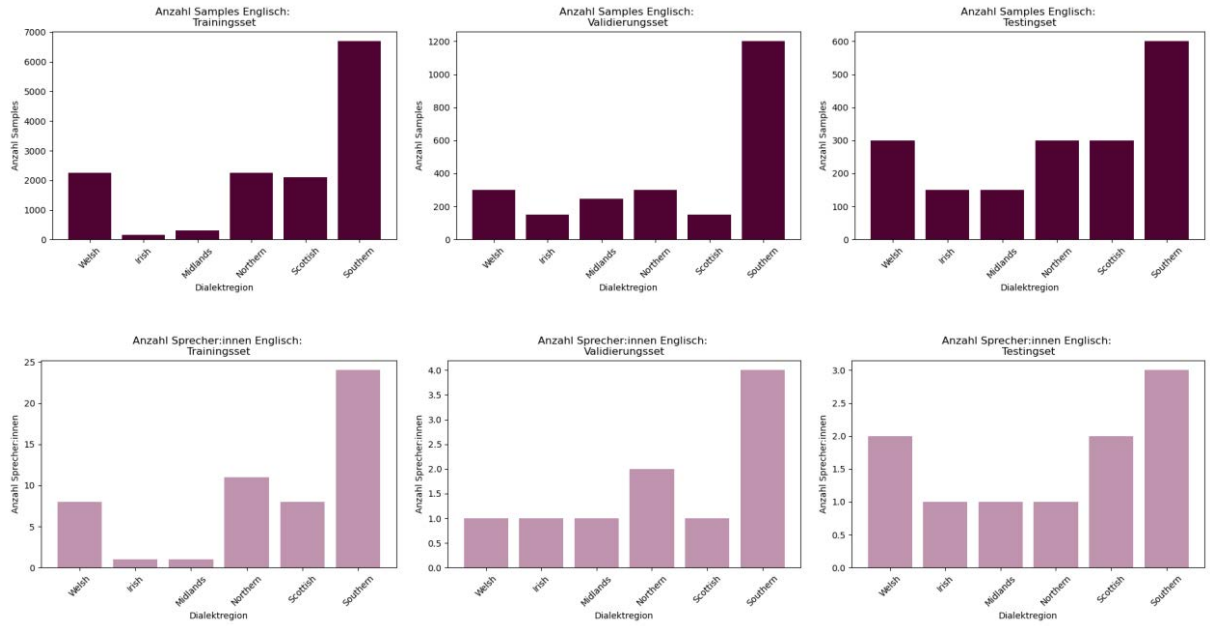


Abbildung 23: Statistiken zu den erstellten Splits für den englischen Korpus

3.3.2 Analyse der Ergebnisse

Die Tabelle 6 zeigt die Ergebnisse des Experiments an den Validierungs- und Testingsets. Insgesamt sind der Macro-F1-Score und die Accuracy äussert hoch für so wenig Daten. Der Unterschied zwischen den einzelnen Regionen ist jedoch auffällig. So wurden Irish und Midlands am Validierungsset und am Testingset schlecht vorhergesagt, während die übrigen Dialekte gut vorhergesagt wurden. Dies ist wahrscheinlich auf die geringe Anzahl an Samples und Sprecher:innen zurückzuführen. Bei Northern und Scottish ist ein grösserer Unterschied zwischen den Validierungs- und Testergebnissen festzustellen. Auch

dies hängt vermutlich mit der geringen Sprecher:innenanzahl zusammen, weil die Ergebnisse bei den Validierungsdaten beziehungsweise Testingdaten abhängig sind von eine:r einzelnen Sprecher:in.

	VALIDIERUNGSSET		TESTINGSET	
CLASS	**f1-score**	**support**	**f1-score**	**support**
WELSH	0.9286	299	0.9396	300
IRISH	0	150	0.1595	150
MIDLANDS	0	246	0.0599	150
NORTHERN	0.4923	300	0.1865	300
SCOTTISH	0.4943	150	0.8914	300
SOUTHERN	0.7718	1200	0.7147	600
ACCURACY	0.6687	-	0.6694	-
MACRO AVG	0.4478	2345	0.4919	1800

Tabelle 6: Ergebnisse des Valididerungs- und Testingsets für das Experiment mit dem englischen Korpus

Die Confusion-Matrizen in den Abbildung 24 widerspiegeln, was bereits die Metriken anzeigen, und unterstreichen die Unterschiede in der Anzahl der Samples zwischen den Regionen. Die Regionen werden gerade mit Southern am häufigsten verwechselt, was damit zusammenhängen könnte, dass das Modell wegen der hohen Sampleanzahl zu einseitig auf diesen Dialekt trainiert wurde. Die Dysbalance in der Anzahl Samples ist hier folglich ein Nachteil. Die Confusion-Matrix zeigt aber, dass die Vorhersagen tendenziell in die richtige Richtung gehen. Bei Scottish, Southern und Welsh wird jeweils der korrekte Dialekt beim Validierungs- und Testingset am meisten vorhergesagt. Bei Northern ist dies zumindest im Validierungsdatenset der Fall. Wobei sich auch hier der erwähnte Effekt manifestiert, dass die Ergebnisse nur von eine:r Sprecher:in abhängen.

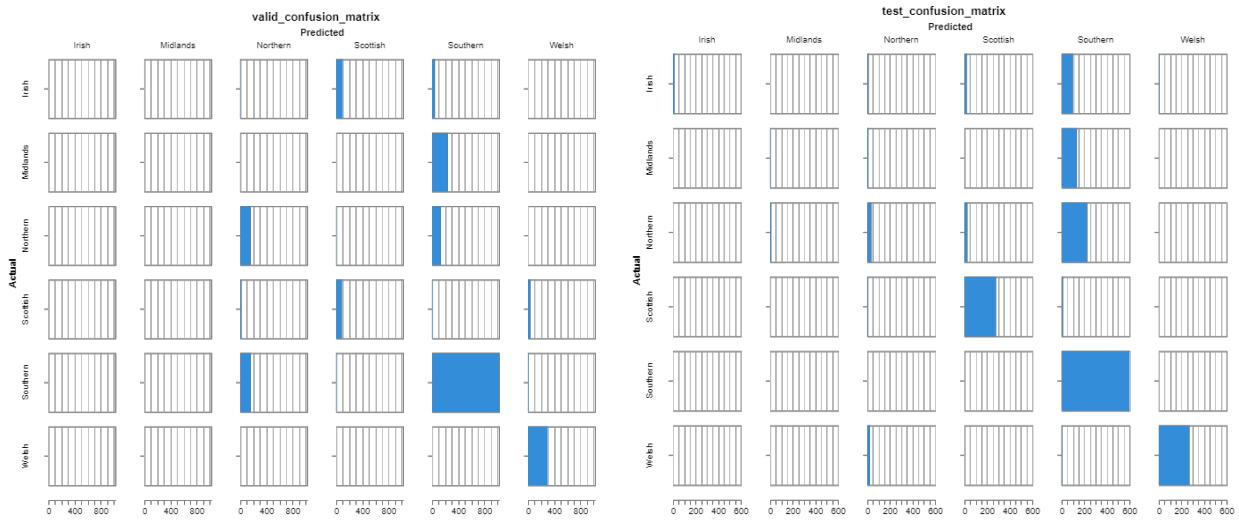


Abbildung 24: Confusion-Matrizen für die Vorhersagen am englischen Korpus

3.3.3 Fazit

Obwohl die geringe Anzahl von Samples und Sprecher:innen eine konkrete Aussage erschwert, deuten die hohen Scores und die Confusion-Matrizen darauf hin, dass das Modell die Dialekte (bis auf Irish und Midlands) bereits gut vorhersagt. Gerade weil so wenig Daten für das Training des Modells vorhanden waren, unterstützen die guten Ergebnisse die These, dass Whisper einen Vorsprung für Englisch mitbringt. Wie ausgeprägt dieser Vorteil ist, würde sich in einem weiteren Experiment näher untersuchen lassen, bei dem zum Beispiel mehr englische Daten hinzugezogen werden, so dass das Modell mit etwa gleich vielen Daten wie bei den schweizerdeutschen Korpora trainiert wird. Weil ein solches Experiment den Rahmen dieser Arbeit sprengen würde, wird im Folgenden stattdessen der STT4SG-350-Korpus so verkleinert, dass er dem englischen Korpus in der Verteilung und Anzahl Samples pro Sprecher:in ähnelt. Ausserdem kann mit einem solchen Experiment weiter untersucht werden, was die ausgeglichene Verteilung der Samples nach Sprecher:innen und die Unausgeglichenheit in der Anzahl Samples pro Dialektregionen für einen Einfluss haben.

3.4 Experiment 4: Ausgeglichene Anzahl Samples pro Sprecher:in

In diesem Experiment wurde wie im Experiment 2: Balancierter STT4SG-350-Korpus der STT4SG-350-Korpus verwendet. Der Korpus gilt, wie in Abschnitt 2.2.2 beschrieben, als ein balancierter Korpus bezüglich der Anzahl Sprecher:innen und Anzahl Samples pro Region. Für dieses Experiment wurde jedoch die Anzahl Samples reduziert, sodass in jeder Dialektregion ungefähr gleich viele Samples pro Sprecher:in vorhanden sind. Dies geschah mit der Absicht, wie in Abschnitt 3.3.3 erwähnt, einen Datensatz zu schaffen, der vergleichbar mit dem Englischen ist und um die Auswirkungen einer ausgeglichenen Verteilung der Samples nach Sprecher:innen und die Unausgeglichenheit in der Anzahl Samples pro Dialektregionen zu untersuchen.

3.4.1 Set-up

Die Anzahl zu erreichender Samples pro Sprecher:in wurde anhand der Verteilung der Samples pro Sprecher:in in der entsprechenden Region festgelegt. Sprecher:innen, deren Anzahl Samples unter dem festgelegten Wert lagen, wurden inklusive ihrer Samples ausgeschlossen. Bei allen übrigen Sprecher:innen wurde die Anzahl Samples auf die vordefinierte Zielanzahl reduziert. Die Herausforderung bestand darin, so wenig Sprecher:innen wie möglich auszuschliessen, da eine möglichst grosse Anzahl Sprecher:innen einen Einfluss auf die Dialekterkennung hat (wie in Abschnitt 3.2.3 erklärt). Gleichzeitig galt es, nicht unnötig Samples zu verlieren. Die Zielanzahl der Samples pro Sprecher:in in einer Region wurden so bestimmt, dass maximal 10 % der Sprecher:innen dieser Region darunterlagen. Aufgrund von Abweichungen in den Verteilungen der Samples pro Sprecher:in zwischen den Regionen variiert die Zielanzahl. Folglich gibt es Abweichungen bzgl. der Gesamtanzahl Samples pro Region. Innerhalb jeder Region hat jedoch jede:r Sprecher:in die gleiche Anzahl Samples. Die Sample- und Sprecher:innen-Statistiken des gesamten Datensets sind in Abbildung 25 zu sehen.

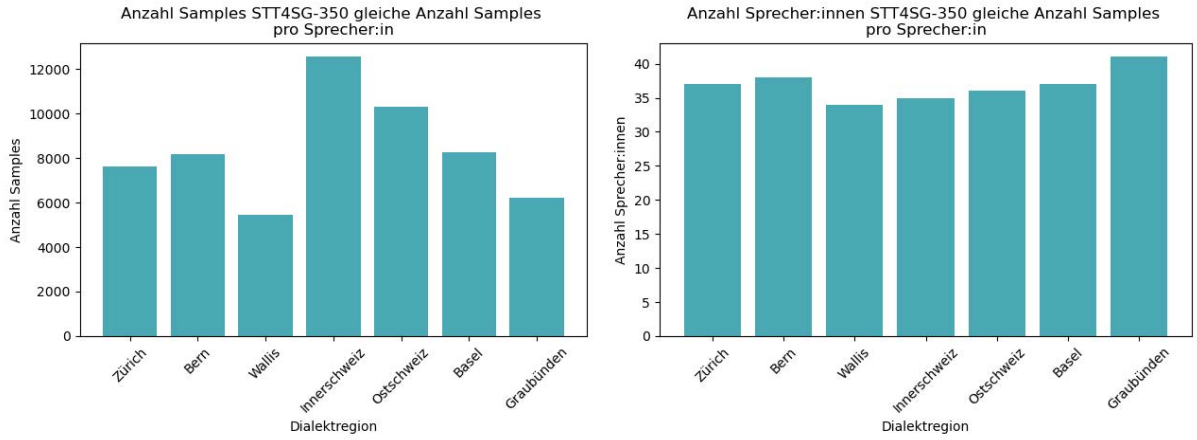


Abbildung 25: Statistiken zum Datenset mit einer ausgeglichenen Anzahl Samples pro Sprecher:in

Der Datensatz wurde dann mithilfe des in Abschnitt 3.1.2 beschriebenen iterativen Verfahrens in Trainings-, Validierungs- und Testingset aufgeteilt. In der Abbildung 26 sind die Sprecher:innen- und Sampleverteilungen der Splits dargestellt.

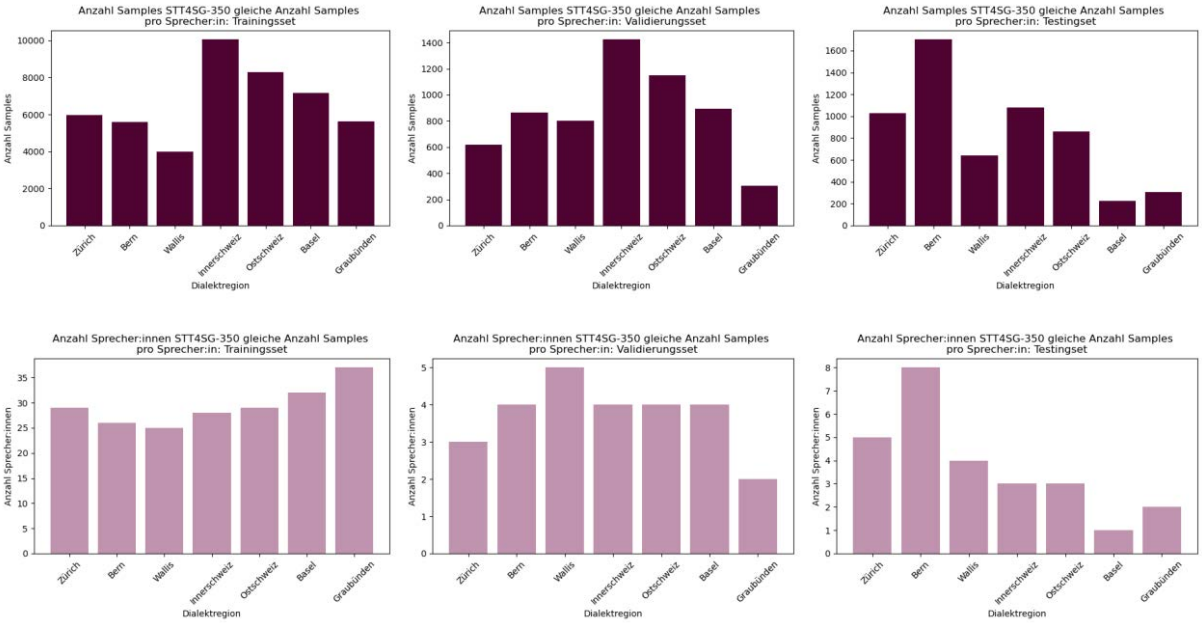


Abbildung 26: Statistiken zu den erstellten Splits für das Datenset mit einer ausgeglichenen Anzahl Samples pro Sprecher:in

3.4.2 Analyse der Ergebnisse

Die Ergebnisse dieses Experiments in der Tabelle 7 unterscheiden sich zwischen den Regionen sowie zwischen den Validierungs- und Testingdaten. Zürich und Graubünden weisen im Validierungsset schlechtere Ergebnisse auf als im Testingset, was wahrscheinlich auf ihre geringe Anzahl Samples und

Sprecher:innen zurückzuführen ist. Im Testingset ist für Zürich der Anteil an korrekt zugewiesenen Samples zwar höher, aber dennoch wird der Grossteil der Samples von Zürich nicht erkannt, wie auch in der Confusion-Matrix in Abbildung 27 zu beobachten ist. Basel hingegen wird im Validierungsset wesentlich besser erkannt als im Testingset. Dies liegt daran, dass Basel nur mit eine:r Sprecher:in im Testingset vertreten ist. Dieser Nachteil wurde schon im Experiment 3: Englischer Korpus mit dem englischen Korpus festgestellt.

	VALIDIERUNGSSET		TESTINGSET	
CLASS	**f1-score**	**support**	**f1-score**	**support**
ZÜRICH	0.1	618	0.1857	1030
INNERSCHWEIZ	0.2082	1423	0.4678	1077
GRAUBÜNDEN	0.1652	304	0.4679	304
BASEL	0.5601	896	0.2390	224
OSTSCHWEIZ	0.4627	1148	0.6252	861
WALLIS	0.5043	800	0.5119	640
BERN	0.5341	864	0.5215	1701
ACCURACY	0.3686	-	0.4675	-
MACRO AVG	0.3621	6053	0.4313	5837

Tabelle 7: Ergebnisse des Valididerungs- und Testingsets für das Experiment mit ausgeglichener Anzahl Samples

Innerschweiz weist in beiden Sets die höchste Anzahl Samples auf, wird allerdings nur im Testingset am meisten vorhergesagt, was in der Confusion-Matrix in Abbildung 27 gut ersichtlich ist. Dadurch, dass die Innerschweiz auch im Trainingsset die höchste Anzahl an Samples aufweist, könnte dies ebenfalls ein Hinweis darauf sein, dass das Modell im Zweifelsfall Samples der Region zuweist, mit der sie am meisten trainiert wurde. Ähnliches wurde schon im Abschnitt 3.3.2 bei der Region Southern beobachtet.

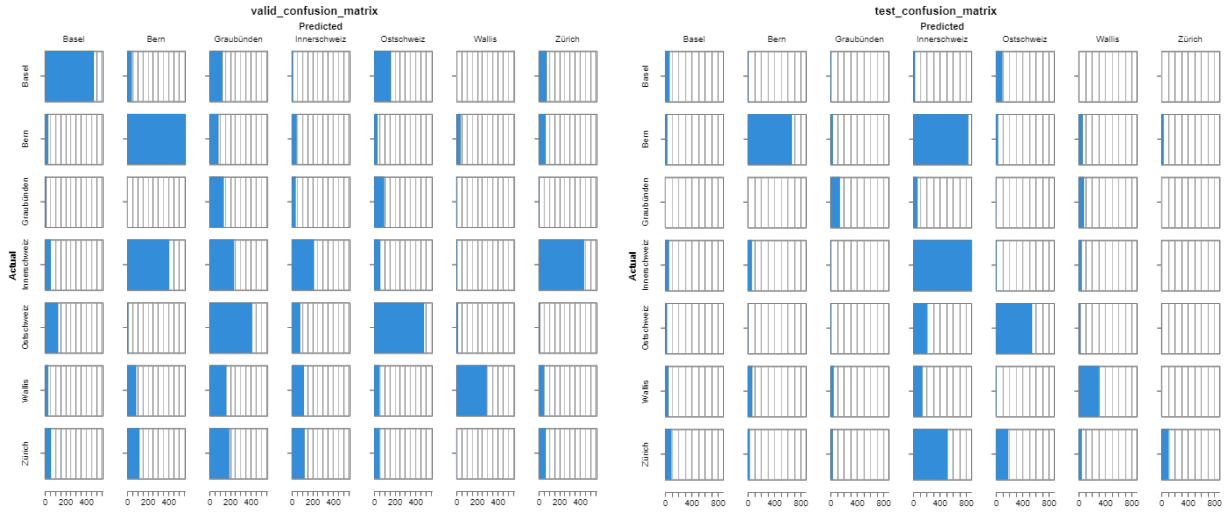


Abbildung 27: Confusion-Matrizen für die Vorhersagen am Datenset mit einer ausgeglichenen Anzahl Samples pro Sprecher:in

Vergleicht man die Ergebnisse mit dem Experiment 2: Balancierter STT4SG-350-Korpus, so fällt die Gesamtleistung des Models sowohl im Validierungsset- als auch im Testingset leicht besser aus. Dies deutet darauf hin, dass die Anzahl Samples pro Sprecher:in Einfluss auf die Leistung hat.

Wie im Abschnitt 3.3.3 erwähnt, erzielten die Ergebnisse des Experiments auf dem englischen Datensatz trotz weniger Samples bessere Scores. Der verwendete Datensatz für dieses Experiment hat zwar eine vergleichbare Verteilung bezüglich Anzahl Samples pro Sprecher:in wie der englische, enthält aber gesamthaft mehr Daten. Obwohl dem so ist, fallen die Ergebnisse im Vergleich zum Experiment 3: Englischer Korpus mit dem englischen Korpus schlechter aus. Dies bestärkt die Annahme aus Abschnitt 3.3.3, dass die englischen Daten von dem Pretraining des Whisper-Modells auf englischen Daten profitieren[10].

3.4.3 Fazit

Interessanterweise performt das Modell in diesem Experiment besser, obwohl im Vergleich zum vollständigen STT4SG-Korpus sowohl die Anzahl Samples als auch die Anzahl Sprecher:innen reduziert wurden. Eine Reduktion der Anzahl Samples pro Sprecher:in, um eine ausgeglichene Verteilung zu erreichen, scheint demnach eine Verbesserung zu bewirken. Hingegen führt die Unausgeglichenheit der Anzahl Samples pro Region ähnlich wie im englischen Datensatz dazu, dass der Region mit übermässig vielen Samples mehr andere Samples zugeordnet werden. Performanceeinbussen des Modells im Testingset waren wie beim englischen Experiment betreffend einer zu geringen Anzahl Sprecher:innen beobachtbar. Weil trotz mehr Samples keine besseren Scores erzielt werden konnten als mit dem englischen Datensatz, ist davon auszugehen, dass Englisch von einem Vorteil wegen des Pretrainings profitiert.

3.5 Experiment 5: Korpus-Mix 30000 Samples pro Dialektregion

Bisher wurde gezeigt, dass sich eine grössere Anzahl Sprecher:innen, eine ausgeglichene Anzahl Samples pro Sprecher:in und eine ausgeglichen Anzahl Samples pro Dialektregion positiv auf die Ergebnisse auswirkt. Noch unklar ist, welchen Einfluss eine grössere Anzahl Samples pro Dialektregion hat. In Experiment 2 wurde das Modell zwar mit einer grösseren Anzahl Samples pro Dialektregion trainiert als in Experiment 1 und lieferte schlechtere Resultate als der auf 10000 Samples pro Region reduzierte gemischte Korpus. Doch waren auch weniger Sprecher:innen vorhanden, so dass unklar ist, ob es nicht daran lag. In einem abschliessenden Experiment wird mit dem optimierten Preprocessing[11], das beim ersten gemischten Korpus noch nicht eingesetzt wurde, ein zweiter, grösserer gemischter, aber immer noch ausgeglichener Korpus erstellt und nochmals ein Modell basierend auf Whisper trainiert. So kann im direkten Vergleich mit dem ersten Experiment bei einer gleichbleibenden Anzahl Sprecher:innen geprüft werden, ob mehr Samples pro Dialektregion einen positiven oder negativen Einfluss auf das Training haben.

[10] Wie es sich mit der Merkmalserkennung von schweizerdeutschen Dialekten verhält, ist nicht Teil dieser Arbeit. Es ist an dieser Stelle jedoch auch nicht auszuschliessen, dass schweizerdeutsche Dialekte näher beieinander liegen als englische und deren Erkennung damit schwieriger sein könnte.

[11] Es ist davon auszugehen, dass das optimierte Preprocessing keinen Einfluss auf das Training oder die Resultate hat.

3.5.1 Set-up

Für einen Korpus mit einer maximalen, aber immer noch ausgeglichen Anzahl Samples pro Dialektregion musste das Datenset so reduziert werden, dass alle Regionen gleich viele Samples haben wie die Region mit den wenigsten Samples. Konkateniert man den STT4SG-350- und der SDS-200-Korpus ist dies Graubünden mit 34463 Samples. Diese Zahl wurde auf 30000 abgerundet, so dass sich die in Abbildung 28 dargestellte Verteilung der Samples und Sprecher:innen ergab. Die Anzahl Sprecher:innen ist die gleiche wie in Experiment 1, da in beiden Experimenten alle vorhandenen Sprecher:innen verwendet werden. Also ist nur die Anzahl Samples um 20000 höher.

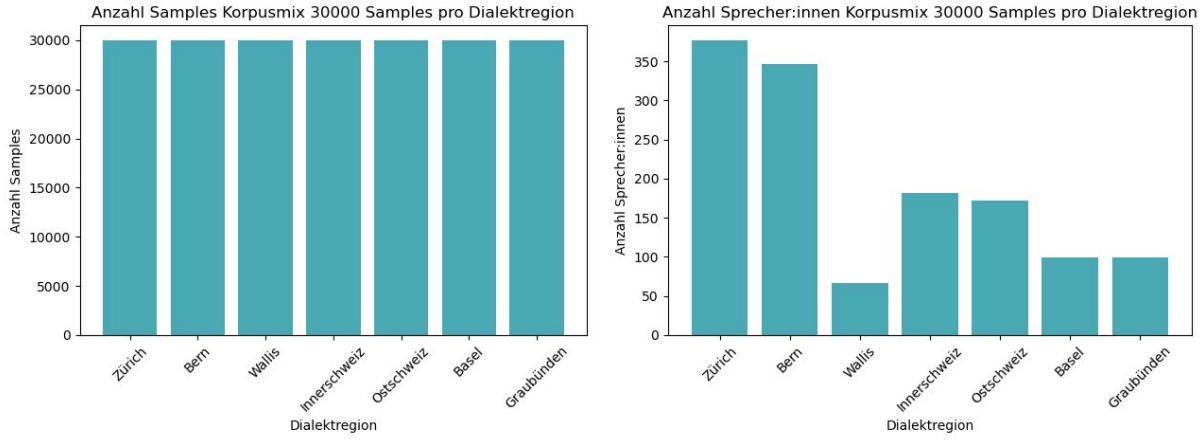


Abbildung 28: Statistik zum Korpusmix mit 30000 Samples pro Dialektregion

Die erstellten Splits sind in Abbildung 29 zu sehen. Leider ist erst im Nachhinein augefallen, dass Basel im Validierungsset zu wenig Samples zugeteilt wurden, worauf in der Analyse geachtet werden muss. Weil das Trainingsset jedoch ausgeglichen ist und die Basel-Daten im Validierungsset nur das Update der Learning-Rate in geringem Mass beeinflusst, sollte das Training nicht oder nur wenig beeinträchtigt worden sein.

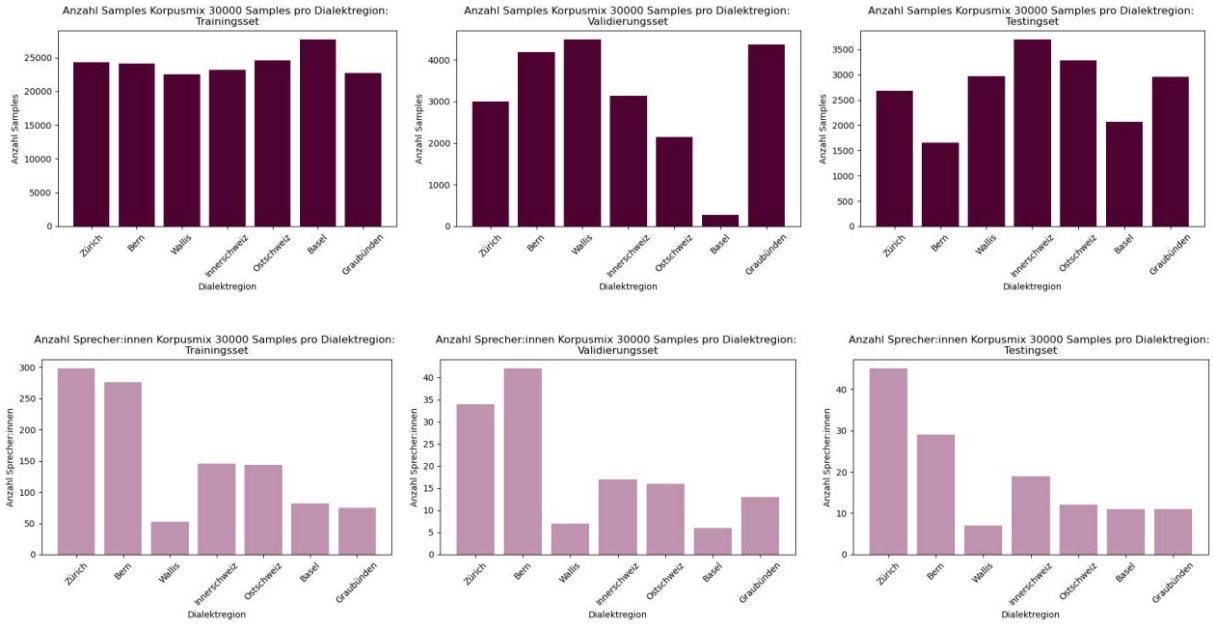


Abbildung 29: Statistiken zu den erstellten Splits für den Korpusmix mit 30000 Samples pro Dialektregion

3.5.2 Der Trainingsprozess (Exkurs)

Bisher wurde nicht speziell auf den Trainingsvorgang eingegangen. Der Fokus dieser Arbeit liegt auf der Reproduktion und Analyse der Ergebnisse und weniger auf der Optimierung des Trainings. Wie aber in der Abbildung 30 besonders an der Entwicklung des Validierungs-Loss und der Validierungs-Accuracy zu sehen ist, ist der Trainingsprozess nicht ideal. Dies wurde auch in anderen Experimenten mehr oder weniger ausgeprägt beobachtet. Während der Loss am Trainingsset kleiner und die Accuracy grösser wird - was zu erwarten wäre -, steigt der Loss am Validierungsset und die Accuracy schwank stark mit einem leicht sinkenden Trend. Das Modell erreicht also gleich anfangs die besten Validierungsresultate und wird dann eher schlechter. Dies deutet auf ein sehr frühes Overfitting hin. Wie eingangs im Abschnitt 2.3.1 erwähnt, wurde in vorhergehenden Arbeiten beobachtet, dass Wav2Vec XLS-R mit nur wenig Daten viel leistet. Zudem ist in zwei Blogs von ähnlichen Overfitting-Erscheinungen beim Finetuning von Whisper zu lesen, wenn nur wenig Trainingsdaten vorhanden sind [46][47]. Es könnte sich folglich um ein bekanntes Phänomen handeln, das aus einer Kombination aus leistungsstarkem Pretrained-Model und wenig Finetuning-Daten resultiert. In folgenden Arbeiten, in denen es darum gehen soll, das Modell zu verbessern, sollte dies nochmals untersucht werden. So könnten Hyperparameter-Tuning und Dropout-Layers Abhilfe schaffen. Denn Ziel wäre, dass sich das Modell während mehr als einer Epoche zuverlässig und stetig verbessert, der Loss sinkt und die Accuracy steigt.

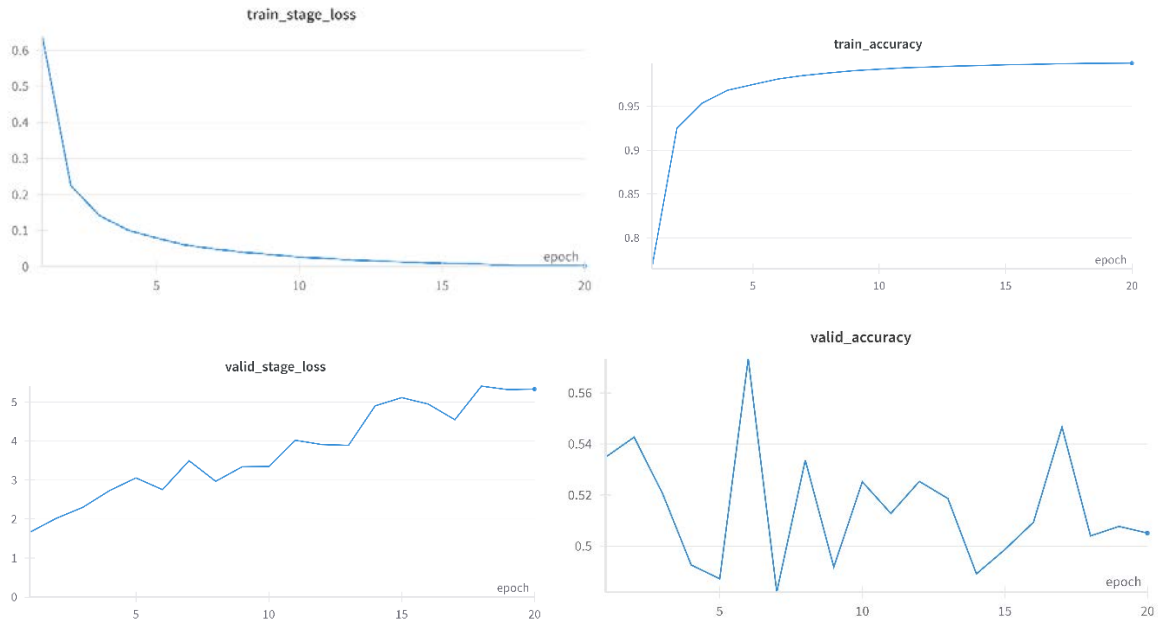


Abbildung 30: Statistiken zum Trainingsprozess des Modells

3.5.3 Analyse der Ergebnisse

In diesem Experiment wurden nahezu gleiche F1-Scores am Testingset erreicht wie im ersten Experiment (siehe Tabelle 9). Die Validierungsscores, wie in Tabelle 8 aufgeführt, variieren stärker. Wie erwartet ist Basel wegen des geringen Supports bei den Validierungsscores in diesem Experiment besonders auffällig und hat einen tiefen F1-Score. Dies wirkt sich auch auf den Micro- und den Macro-F1-Score aus. Insgesamt schneidet dieses Experiment jedoch etwas schlechter ab. Es ist zu beachten, dass die höhere Anzahl Samples im Datenset bedeutet, dass auch der Support (bis auf Basel im Validierungsset) bei der Validierung und dem Testing grösser ist und die Scores dadurch repräsentativer sind für die Performance des Modells.

Dass die Innerschweiz am schlechtesten vorhergesagt wird, wird bei diesem Experiment noch deutlicher. Folglich hat das bestehende Modell unabhängig von der Datenmenge, mit der es trainiert wurde, Schwierigkeiten, Samples der Innerschweiz zu identifizieren. Auch Bern schneidet schlechter ab. Hingegen werden Zürich, Basel und die Ostschweiz besser vorhergesagt. Bei den anderen Dialekten besteht also ein gewisser Spielraum nach oben und unten. Aber ohne genauere Analyse und weitere Experimente lässt sich nicht sagen, ob die Unterschiede zwischen den Scores der Experimente zufällig sind oder tatsächlich von der Anzahl Samples abhängen.

VALIDIERUNGSSCORES

DATENSET	10000 Samples pro Dialektregion		30000 Samples pro Dialektregion	
KLASSE	F1-Score	Support	F1-Score	Support
ZÜRICH	0.5968	908	0.4089	3005
INNERSCHWEIZ	0.4670	1001	0.3440	3133
GRAUBÜNDEN	0.7411	1344	0.5896	4371
BASEL	0.6689	1042	0.1667	271
OSTSCHWEIZ	0.6321	907	0.5034	2142
WALLIS	0.8988	1094	0.5655	4485
BERN	0.7200	882	0.6859	4182
ACCURACY	**0.6807**	-	**0.5051**	-
MACRO AVG	**0.6750**	7178	**0.4663**	21589

Tabelle 8: Ergebnisse für das Validierungsset zu den Experimenten 1 und 5

TESTINGSCORES

DATENSET	10000 Samples pro Dialektregion		30000 Samples pro Dialektregion	
KLASSE	F1-Score	Support	F1-Score	Support
ZÜRICH	0.5227	862	0.5495	2687
INNERSCHWEIZ	0.4429	1277	0.3656	3694
GRAUBÜNDEN	0.3439	681	0.3708	2954
BASEL	0.4936	882	0.6541	2072
OSTSCHWEIZ	0.6793	1116	0.7546	3282
WALLIS	0.6983	969	0.6658	2974
BERN	0.5631	893	0.4098	1657
ACCURACY	**0.5445**	-	**0.5395**	-
MACRO AVG	**0.5348**	6680	**0.5386**	19320

Tabelle 9: Ergebnisse zum Testingset den Experimenten 1 und 5

In der Confusion-Matrix in Abbildung 31 tritt die Verwechslungsgefahr mit Bern und der Innerschweiz verglichen mit dem kleineren gemischten Korpus stärker hervor, während die Verwechslungsgefahr mit Zürich weniger ausgeprägt ist. Die Verwechslungstendenzen sind jedoch bei beiden Experimenten ähnlich.

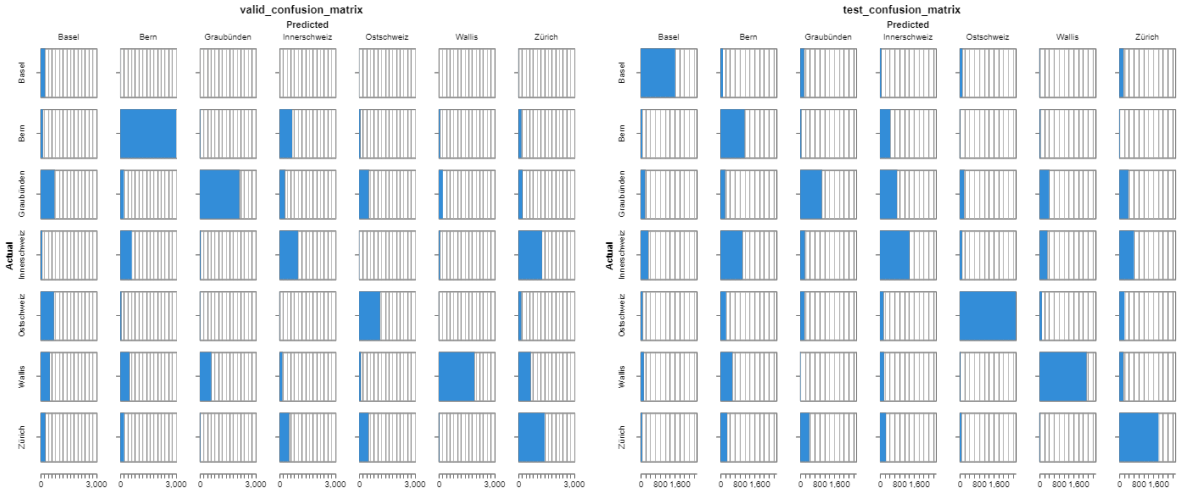


Abbildung 31: Confusion-Matrizen zu den Vorhersagen für den Korpusmix mit 30000 Samples pro Dialektregion

In den Heatmaps, die darstellen, wie klar die Samples der jeweilige Sprecher:innen zugeordnet wurden, sind grössere Unterschiede zum ersten Experiment zu erkennen. Abbildung 32 zeigt, dass die Sprecher:innen bei der Innerschweiz, Zürich, Graubünden und dem Wallis weniger klar einer oder nur zwei bis drei Dialektregionen zugeordet werden.

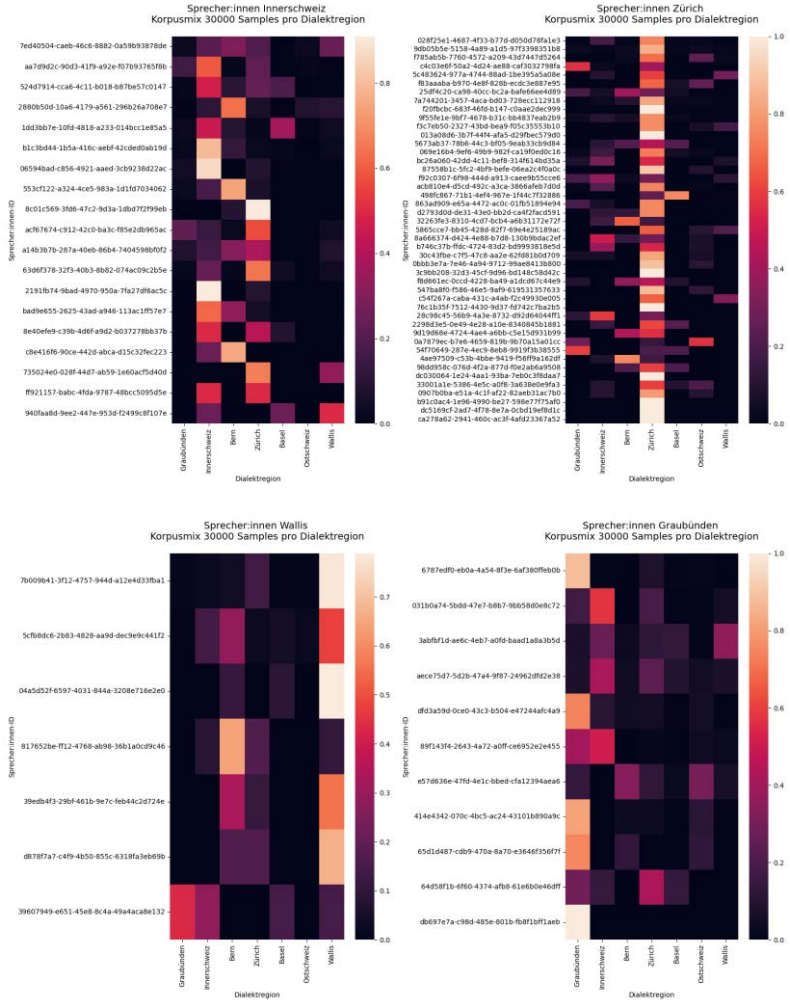


Abbildung 32: Heatmaps für die Zuordnung der Samples nach Sprecher:innen für die Innerschweiz, Zürich, das Wallis und Graubünden

Die Sprecher:innen aus Basel und Bern werden umgekehrt klarer zugeordnet als zuvor, wie in Abbildung 33 ersichtlich. (Die hier nicht dargestellte Heatmap der Ostschweiz gleicht der des ersten Experiments.)

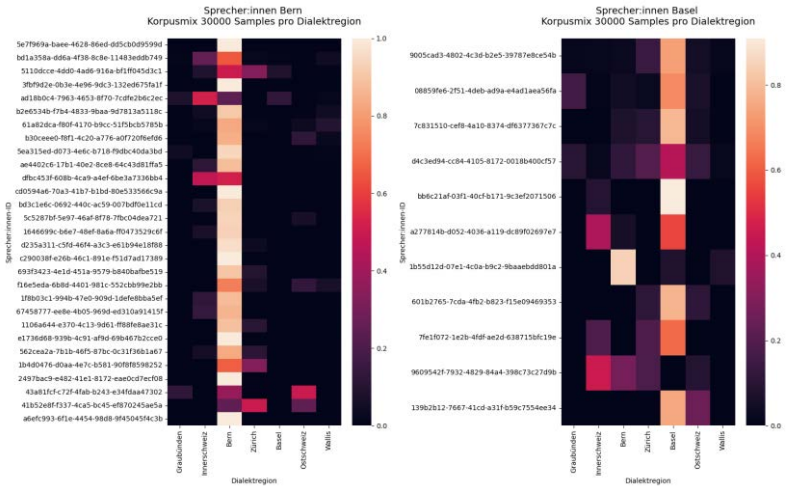


Abbildung 33: Heatmaps für die Zuordnung der Samples nach Sprecher:innen für Bern und Basel

Im Gegensatz zur Confusion-Matrix und den Scores scheint offensichtlicher, dass eine grössere Anzahl Samples pro Sprecher:in beziehungsweise Dialektregion nachteilig sein könnte. Im Zusammenhang mit einem Experiment, in dem die Samples gleicher Sprecher:innen auf die Splits verteilt wurden, haben Frei und Schneider bereits etwas Ähnliches beobachtet. Mit bis zu 10000 Samples pro Dialektregion hätten sich die Ergebnisse verbessert und mit mehr Samples wieder verschlechtert [38]. In späteren Arbeiten müsste allenfalls nochmals systematisch überprüft werden, ob in einem gemischten Korpus ohne überschneidende Sprecher:innen auch so eine Grenze besteht und wo diese liegt. Wie erwähnt, haben mehr Samples allerdings einen möglichen Vorteil: Die Validierungs- und Testingergebnisse sind repräsentativer, weil auch mehr Samples getestet werden. Um diesen Vorteil nutzen zu können, könnte man das Validierungs- und Testingset zum Beispiel mit sonst nicht benötigten Samples der enthaltenen Sprecher:innen auffüllen. Es muss aber darauf geachtet werden, dass das Verhältnis der Anzahl Sprecher:innen über die Splits bestehen bleibt, da bereits gezeigt wurde, dass mehr Sprecher:innen das Resultat verbessern. Hier scheinen Frei und Schneider eine ähnliche Idee angewendet zu haben, wenn auch nicht explizit erklärt. Denn beim Trainingsset ist jeweils «MIX-10000» und beim Validierungsset «MIX-20000» angegeben, was impliziert, dass sie für das Training ein Datenset mit 10000 Samples pro Dialektregion verwendet haben und für die Validierung ein Datenset mit 20000 Samples [38].

Überprüft man wie im ersten Experiment, wo sich die Samples respektive Sprecher:innen geographisch befinden und wie sie zugeordnet wurden, lässt sich das Gleiche beobachten. Bei Dialektregionen, die gut vorhergesagt werden, sind die am besten vorhergesagten Samples im geographischen Zentrum der Region und schlecher vorhergesagte eher am Rand, wie in Abbildung 34 an Beispielen zu erkennen ist. Weitere Beispiele finden sich im Anhang.

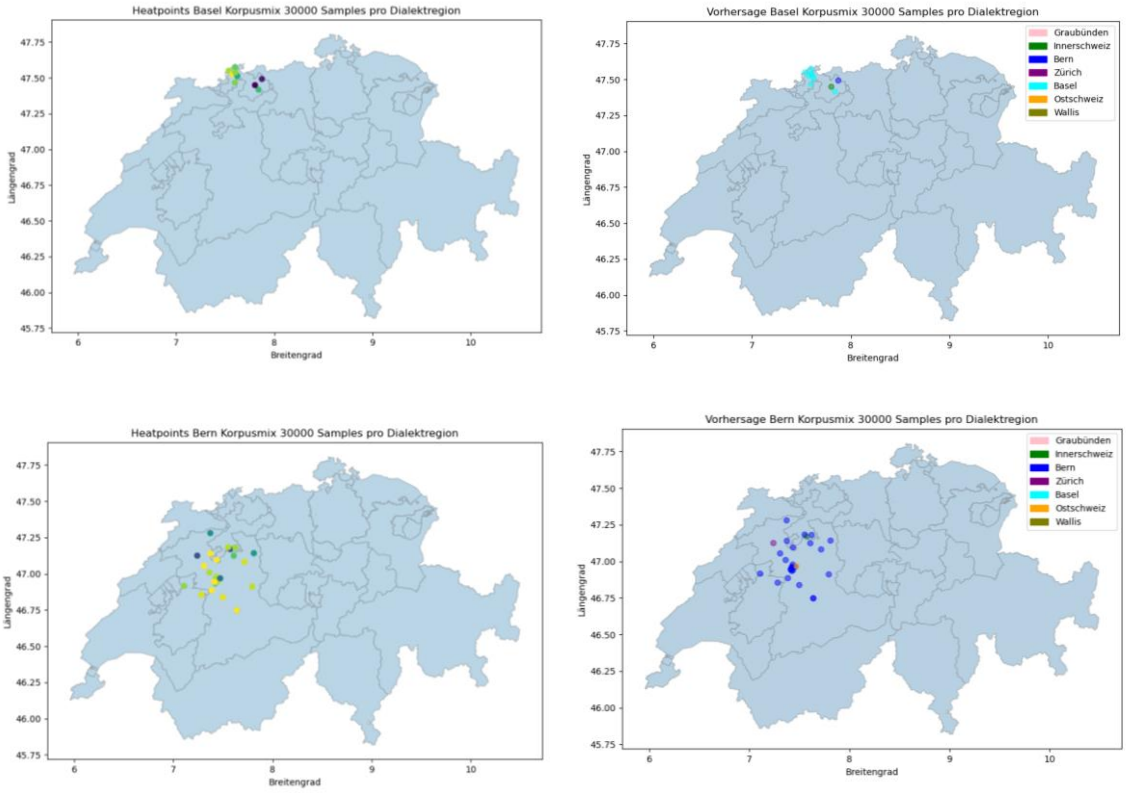


Abbildung 34: Heatpoints und vorhergesagte Labels für Bern und Basel

3.5.4 Fazit

Das abschliessende Experiment führt vor Augen, dass mehr Samples nicht unbedingt bessere Resultate bedeuten. Vielmehr scheint es eine gewisse Schwelle zu geben, bei der eine ideale Anzahl Samples im Verhältnis zur Anzahl Sprecher:innen erreicht wird und ab der sich die Resultate wieder verschlechtern. Wo diese Schwelle genau liegt, sollte in weiteren Arbeiten überprüft werden. Die unterschiedlich guten Resultate von verschiedenen Regionen bei mehr oder weniger Sampels könnte auch ein Hinweis darauf sein, dass diese Grenze nicht bei allen Regionen gleich ist. Ansonsten wurden einige Beobachtungen vor allem des ersten Experiments nochmals bestätigt: Ein gemischter Korpus liefert grundsätzlich die besten Ergebnisse, mehr Sprecher:innen scheinen vorteilhaft, die Innerschweiz wird am schlechtesten erkannt und die geographische Lage der Samples scheint in einem gewissen Masse zu beeinflussen, wie gut das Modell den Dialekt erkennt.

4. Diskussion und Ausblick

Anhand der Experimente zur Dialekterkennung des Schweizerdeutschen mit einem einfachen, auf Whisper basierenden Modell konnten einige architekturunabhängige Faktoren evaluiert werden, die die Resultate des Modells positiv oder negativ beeinflussen. Ein Schwerpunkt lag auf der Zusammenstellung der Trainingsdaten. So verbessern eine möglichst grosse Anzahl Sprecher:innen und eine möglichst ausgeglichen Anzahl Samples pro Dialektregion und Sprecher:in die Ergebnisse. Gleichzeitig heisst «mehr Daten» nicht unbedingt «bessere Ergebnisse». Dies erscheint zunächst kontraintuitiv, gerade weil Schweizerdeutsch eine Low-Resource-Sprache ist. Doch die Experimente führen vor Augen, dass das Verhältnis der Anzahl Samples zur Anzahl Sprecher:innen entscheidend ist und es vermutlich, wie beschrieben eine obere Grenze für alle Dialektregionen zusammen oder pro Dialektregion gibt, bis zu der sich die Resultate mit ansteigender Anzahl Samples verbessern und sich wieder verschlechtern, wenn diese überschritten wird. Nachfolgende Arbeiten sollten dies nochmals untersuchen und diese Grenze(n), falls tatsächlich vorhanden, ermitteln.

Ein weiterer negativer Faktor ist die Tonqualität. Stichproben haben gezeigt, dass Audiodateien mit Nebengeräuschen oder undeutlich Gesprochenes schlecht erkannt wurden, was auch dem Training schaden könnte. Für weitere Experimente könnte es sich also lohnen, solche Samples automatisch mit Noise-Detection-Verfahren zu filtern und Nebengeräusche maschinell zu entfernen.

Dank dem Vergleich mit dem englischen Korpus hat sich zudem gezeigt, dass das Pretraining von Whisper auf bis zu zwei Dritteln englischen Daten einen Vorteil für das Finetuning für englische Dialekte sein könnte. Dies bedeutet umgekehrt, dass beim Finetuning der Schweizerdeutschen Dialekte wahrscheinlich architektonisch und/oder datentechnisch mehr geleistet werden muss als für Englisch, um diesen Vorsprung auszugleichen. Um die Vermutung zu bestätigen, könnte mit grösseren englischen Korpora experimentiert werden.

Neben diesen Faktoren wurden einige allgemeine Schwierigkeiten bei der Dialekterkennung festgestellt, die von den Dialekteigenschaften ausgehen könnten. Erstens werden nicht alle Dialektregionen gleich gut vorhergesagt. Besonders die Innerschweiz kann das Modell nicht gut identifizieren und verwechselt die Samples vor allem mit Bern und Zürich, wenn diese im Datenset sehr präsent sind. Umgekehrt wird die Innerschweiz gut erkannt, wenn Bern und Zürich weniger präsent sind, während Bern und Zürich mit der Innerschweiz verwechselt werden. Diese drei Regionen können also besonders schlecht auseinandergehalten werden. Ob dies wirklich keine modellspezifische Schwäche ist, sondern etwas mit den Dialekten zu tun hat, wäre noch an weiteren Experimenten zu überprüfen.

Zweitens werden Samples von Sprecher:innen, die mehr im Zentrum einer Dialektregion beheimatet sind, korrekt zugewiesen, während Samples von Sprecher:innen am Rand der Regionen schlechter identifiziert werden. Teilweise wird stattdessen die Nachbarregion erkannt. Zusätzlich werden die Samples bei einzelnen Sprecher:innen nicht nur einer, sondern zu fast gleichen Teilen zwei bis maximal drei Dialektregionen zugeordnet. Aus sprachlicher Sicht ist davon auszugehen, dass einige Sprecher:innen ein Dialektgemisch zum Beispiel von benachbarten Regionen sprechen und unklar ist, zu welchem Dialekt deren Samples gezählt werden sollen. Die beschriebenen Phänomene könnten also damit zusammenhängen, dass das Modell Merkmale von verschiedenen Dialekten erkennt und die Samples

deshalb nicht gemäss der Selbsteinschätzung der Sprecher:innen identifiziert. Ob das Modell jedoch ambige Dialektmerkmale bereits so gut erkennt, sei dahingestellt. Falls die Phänomene in zukünftigen Experimenten mit verbesserten Modellen weiterhin auftreten, könnten diese nochmals untersucht werden.

Zusammengefasst geben die aufgezählten Erkenntnisse eine gute Grundlage für weitere Arbeiten. Nächste Schritte könnten sein, ein auf die erwähnten Faktoren optimiertes Datensets zusammenzustellen, das Modell einem Finetuning zu unterziehen, andere Architekturen zu testen und die Dialekterkennungsleistung zu verbessern. Zudem könnten die beschriebenen Schwierigkeiten weiter untersucht und analysiert werden, welche Sprachmerkmale das Modell lernt.

Dankesworte:

Als Erstes danken wir unserer Betreuerin Jasmina und unserem Betreuer Mark für die zielführende Begleitung und den gewährten Freiraum bei der Ausgestaltung der Arbeit. Wir schätzen die konstruktiven Inputs in den verschiedenen Phasen dieser Arbeit und die Expertise bezüglich der Analyse unserer Experimente. Ein ebenso grosser Dank gilt zudem unserem engsten Umfeld, das uns auf unserem Weg stets mit Zuversicht, Ermutigung und viel Geduld unterstützt hat.

5. Verzeichnisse

5.1 Literaturverzeichnis

[1] J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, Automatic Speech Recognition. A Bridge to Practical Applications. 1. Auflage Waltham, MA: Elsevier, 2015, S. 3-50

[2] A. Steigerwald. (4.3.2021). *Automatische Spracherkennung – was sie für Kommunikation und Marketing leisten kann und welche Schweizer Dialekte sie (besser) versteht* [Online]. URL: https://hub.hslu.ch/ikm/2021/03/04/automatische-spracherkennung-in-kommunikation-und-marketing/ [Stand: 16.12.2023]

[3] M. Khan, U. Kifayat, Y. Alharbi, A. Alferaidi, T. S. Alharbi, K. Yadav, N. Alsharabi, A. Ahmad. (25.7.2023). *Understanding the Research Challenges in Low-Resource Language and Linking Bilingual News Articles in Multilingual News Archive* (*Applied Sciences* 13, no. 15: 8566) [Online]. URL: https://doi.org/10.3390/app13158566 [Stand: 17.12.23]

[4] M. Hutchinson. (27.4.2023). *How to Overcome the Need for Data for Low-Resource Languages* [Online]. URL: https://vistatec.com/de/how-to-overcome-the-need-for-data-for-low-resource-languages/ [Stand: 17.12.2023]

[5] G. Lample, A. Conneau. (22.9.2019). *Cross-lingual Language Model Pretraining* [Online]. URL: https://arxiv.org/pdf/1901.07291.pdf [Stand: 17.12.2023]

[6] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J. Wen, J. Yuan, W. X. Zhao, J. Zhu. (26.8.2021) *Pre-trained models: Past, present and future*. URL: https://doi.org/10.1016/j.aiopen.2021.08.002 [Online]. [Stand: 23.10.2023]

[7] H. Wang, J. Li, H. Wu, E. Hovy, Y. Sun. (7.9.2022). *Pre-Trained Language Models and Their Applications* [Online]. URL: https://www.sciencedirect.com/science/article/pii/S2095809922006324 [Stand: 20.12.2023]

[8] M. Phi. (1.5.2020). *Illustrated Guide to Transformers- Step by Step Explanation* [Online]. https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0 [Stand: 14.11.2023]

[9] R. Krüger. (2021). *Die Transformer-Architektur für Systeme zur neuronalen maschinellen Übersetzung – eine popularisierende Darstellung* (Wissenschaftliche Zeitschrift für Translation und Fachkommunikation trans-kom 14 [2] (2021): 278–324) [Online]. URL: https://www.trans-kom.eu/bd14nr02/trans-kom_14_02_05_Krueger_NMUe.20211202.pdf [Stand: 23.10.2023]

[10] (o.N.) (8.11.2023). Transformer model in NLP : Your AI and ML questions, answered [Online]. URL : https://www.capitalone.com/tech/machine-learning/transformer-nlp/ [Stand: 14.12.2023]

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. (2017) *Attention Is All You Need* [Online]. URL: https://arxiv.org/abs/1706.03762 [Stand: 1.12.2023]

[12] A. Baevksi, H. Zhou, A. Mohamed, M. Auli. (22.10.2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* [Online]. URL: https://arxiv.org/abs/2006.11477 [Stand: 28.10.2023]

[13] S. Schneider, A. Baevski, R. Collobert, M. Auli. (11.9.2019). wav2vec: Unsupervised Pre-Training for Speech Recognition [Online]. URL: https://arxiv.org/abs/1904.05862 [Stand: 28.10.2023]

[14] A. Babu, C.Wang, A. Tjandra, K. Lukhotia, Q. Xu, N. K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, M. Auli. (16.12.2021). *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale* [Online]. URL: https://arxiv.org/abs/2111.09296 [Stand: 10.12.2023]

[15] A. Radford, J.W. Kim, T. Xu, G.Brockman, C. McLeavey, I. Sutskever. (6.12.2022). *Robust Speech Recognition via Large-Scale Weak Supervision* [Online]. URL: https://arxiv.org/abs/2212.04356 [Stand: 29.10.2023]

[16] J.C. (o.D.). *Whisper API FAQ. General questions about the speech to text API* [Online]. URL: https://help.openai.com/en/articles/7031512-whisper-api-faq# [Stand: 19.12.2023]

[17] (o.N.) (21.9.2022). *Introducing Whisper* [Online]. URL: https://openai.com/research/whisper [Stand: 11.11.2023]

[18] M. Plüss, L. Neukom, C. Scheller, M. Vogel. (8.6.2021). *Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus* [Online]. URL: https://arxiv.org/abs/2010.02810 [Stand: 10.12.2023]

[19] T. Samardžić, Y. Scherrer, E. Glaser. (2016). *ArchiMob - A Corpus of Spoken Swiss German* [Online]. URL: https://aclanthology.org/L16-1641/ [Stand: 10.12.23]

[20] M. Plüss, M. Hürlimann, M.Cuny, A. Stöckli, N. Kapotis, J. Hartmann, M. A. Ulasik, C. Scheller, Y. Schraner, A. Jain, J. Deriu, M. Cieliebak, M. Vogel. (19.5.2022). *SDS-200: A Swiss German Speech to Standard German Text Corpus* [Online]. URL: https://arxiv.org/abs/2205.09501 [Stand: 10.12.2023]

[21] M. Plüss, J. Deriu, Y. Schraner, C. Paonessa, J. Hartmann, L. Schmidt, C. Scheller, M. Hürlimann, T. Samardžić, M. Vogel, M. Cielebak. (30.5.2023). *STT4SG-350: A Speech Corpus for All Swiss German Dialect Regions* [Online]. URL: https://arxiv.org/abs/2305.18855 [Stand: 10.12.2023]

[22] I. Demirşahin, O. Kjartansson, A. Gutkin, C. Rivera. (1.5.2020). *Opensource Multispeaker Corpora of the English Accents in the British Isles* [Online]. URL: https://aclanthology.org/2020.lrec-1.804/ [Stand: 10.12.2023]

[23] (22.4.2021) *The Origins of British Accents* [Online]. URL: https://ndla.no/subject:1:c8d6ed8b-d376-4c7b-b73a-3a1d48c3a357/topic:59a2daf8-db7f-4f47-8160-551f9d9c582c/resource:e6f6b746-fc11-4d0c-b058-a807aaf1eb43 [Stand: 11.11.2023]

[24] G. Mingliang, X. Yuguo. (8.8.2008). *Chinese dialect identification using clustered support vector machine* [Online]. URL: https://ieeexplore.ieee.org/document/4590380 [Stand: 12.12.2023]

[25] L. Nour-Eddine, A. Abdelkader. (2015). *GMM-based maghreb dialect identification system* [Online]. URL: http://dx.doi.org/10.3745/JIPS.02.0015 [Stand: 10.12.2023]

[26] H.C.S. Bougrine, A.Abdelali. (2018). *Spoken arabic algerian dialect identification* [Online] URL: https://www.researchgate.net/publication/325641099_Spoken_Arabic_Algerian_dialect_identification/link/5c64154f92851c48a9d1235e/download [Stand: 10.12.2023]

[27] R. Imaizum, R. Masumura, S. Shiota, H. Kiya. (29.3.2022). *End-to-end Japanese Multi-dialect Speech Recognition and Dialect Identification with Multi-task Learning* [Online]. URL: http://dx.doi.org/10.1561/116.00000045 [Stand: 11.12.23]

[28] M. Alrehaili, T. Alasmari, A. Aoalshutayri. (3.4.2023). *Arabic Speech Dialect Classification using Deep Learning* [Online]. URL: https://ieeexplore.ieee.org/document/10085647 [Stand: 12.12.2023]

[29] (o.N.). (09.08.22). *VarDial 2022*. URL: https://sites.google.com/view/vardial-2022/home [Stand: 12.12.2023]

[30] M. Zampieri, S. Malmasi, P. Nakov, A. Ali, S. Shon, J. Glass, Y. Scherrer, T. Samardžić, N. Ljubešić, J. Tiedemann, C. van der Lee, S. Grondelaers, N. Oostdijk, D. Speelman, A. van den Bosch, R. Kumar. B. Lahiri, M. Jain (20.8.2018). *Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign* [Online]. URL: https://aclanthology.org/W18-3901.pdf [Stand: 12.12.2023]

[31] T. Jauhiainen, H. Jauhiainen K. Lindén (20.8.2018). *HeLI-based Experiments in Swiss German Dialect Identification* [Online]. URL: https://aclanthology.org/W18-3929.pdf [Stand: 12.12.2023]

[32] M. Zampieri, S. Malmasi, Y. Scherrer, T. Samardžić, F. Tyers, M. Silfverberg, N.Klyueva, T. Pan, C. Huang, R.T. Ionescu, A. Burnaru, T. Jauhiainen. (7.7.2019). *Report on the Third VarDial Evaluation Campaign* [Online]. URL: https://aclanthology.org/W19-1401.pdf [Stand: 12.12.2023]

[33] C. Sicard, V. Gillioz, K. Pyszkowski. (13.9.2023). *Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects* [Online]. URL: https://arxiv.org/pdf/2304.11075.pdf [Stand: 10.12.202]

[34] C. Paonessa, Y. Schraner, J. Deriu, M. Hürlimann, M. Vogel, M. Cielebak. (13.10.23). *Dialect Transfer for Swiss German Speech Translation* [Online]. URL: https://arxiv.org/pdf/2310.09088.pdf [Stand: 10.12.2023]

[35] P. Fivian, D. Reiser. (11.6.2021). *Speech Classification using wav2vec 2.0* [Online]. URL: *https://www.zhaw.ch/storage/engineering/institute-zentren/cai/BA21_Speech_Classification_Reiser_Fivian.pdf* [Stand: 10.12.2023]

[36] S. Stucki, P. Randjelovic. (24.12.2021). *Automatic Detection of Swiss German Dialects using Wav2Vec* [Online]. https://www.zhaw.ch/storage/engineering/institute-zentren/cai/Project_Thesis_Automatic_Dialect_Detection.pdf [Stand: 10.12.2023]

[37] S. Stucki, P. Randjelovic. (17.6.2022). *Exploring Wav2Vec2 Pre-Training on Swiss German Dialects using Speech Translation and Classification* [Online]. URL: https://www.zhaw.ch/storage/engineering/institute-zentren/cai/studentische_arbeiten/BA22_ciel_Stucki_Ranjelovic_Wave2Vec_for_Swiss_German.pdf [Stand: 10.12.2023]

[38] C. Frei, P. Schneider. (8.6.2023). *Automatic Identification of Swiss German Dialects using Large Language Models* [Online]. URL: https://www.zhaw.ch/storage/engineering/institute-

zentren/cai/studentische_arbeiten/Spring_2023/Spring23_BA_ciel_Dialect_Recognition_Swiss_German_Schneider_Frei.pdf [Stand: 10.12.2023]

[39] *Speech Brain.* URL: https://speechbrain.github.io/ [Stand: 15.12.2023]

[40] S. Wang, Y. Yang, Y. Qian, K. Yu. (1.3.2021). *Revisiting the Statistics Pooling Layer in Deep Speaker Embedding Learning* [Online]. URL: https://ieeexplore.ieee.org/document/9362097 [Stand: 12.12.2023]

[41] B. Basnet. (27.2.2020). *Log Softmax Vs Softmax* [Online]. URL: https://deepdatascience.wordpress.com/2020/02/27/log-softmax-vs-softmax/ [Stand: 16.12.23]

[42] V.S. Abhirami (10.10.2021). Softmax vs. LogSoftmax [Online]. URL: https://medium.com/@AbhiramiVS/softmax-vs-logsoftmax-eb94254445a2 [Stand: 16.12.23]

[43] (o.N.). (13.5.2021). *What is the advantage of using log softmax instead of softmax?* [Online]. URL:https://datascience.stackexchange.com/questions/40714/what-is-the-advantage-of-using-log-softmax-instead-of-softmax [Stand: 16.12.23]

[44] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le. (3.12.2019). *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition* [Online]. URL: https://arxiv.org/abs/1904.08779 [Stand: 12.12.23]

[45] *Scikit-learn.* URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html [Stand: 16.12.23]

[46] S. Yonash. (14.2.2023). *Automatic Speech Recognition With Whisper. A look at decoding, spelling mistakes and hallucinations, fine-tuning, and more* [Online]. URL: https://betterprogramming.pub/automatic-speech-recognition-with-whisper-84ad29d3b0bd [Stand: 16.12.23]

[47] S. Grundmann. (19.7.2023). *Fine-tuning Whisper for Dutch Language: The Crucial Role of Size* [Online]. URL: https://blog.ml6.eu/fine-tuning-whisper-for-dutch-language-the-crucial-role-of-size-dd5a7012d45f [Stand: 16.12.23]

5.2 Abbildungsverzeichnis

5.3 Tabellenverzeichnis

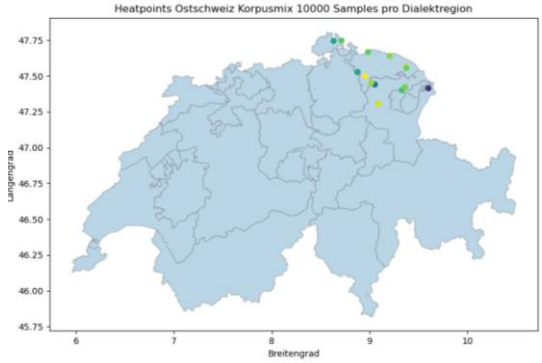
6. Anhang

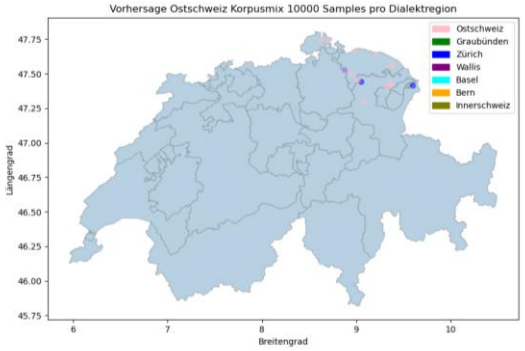
6.1 Quellcode und technische Dokumentation

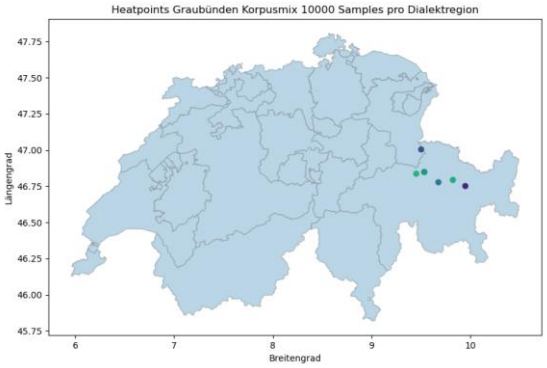
Quellcode und technische Dokumentation finden sich unter folgendem Link:
https://github.zhaw.ch/bollilau/DialectRecognitionPA

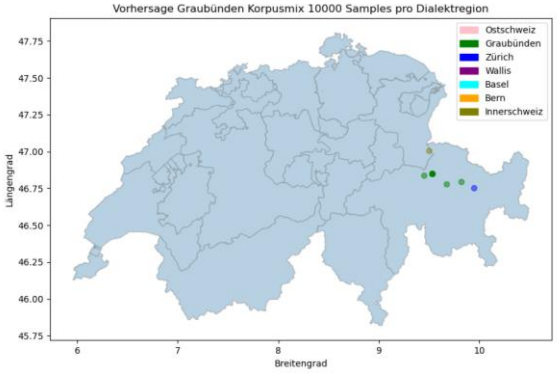
6.2 Ergänzende Diagramme

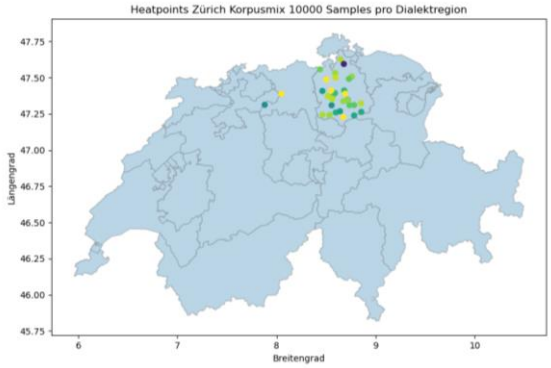
6.2.1 Experiment 1

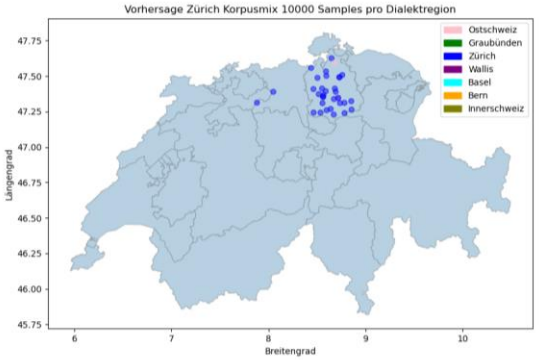


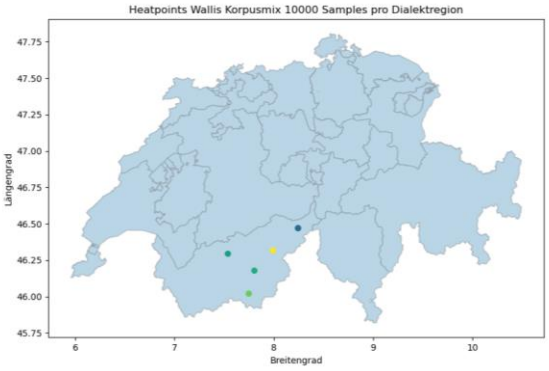


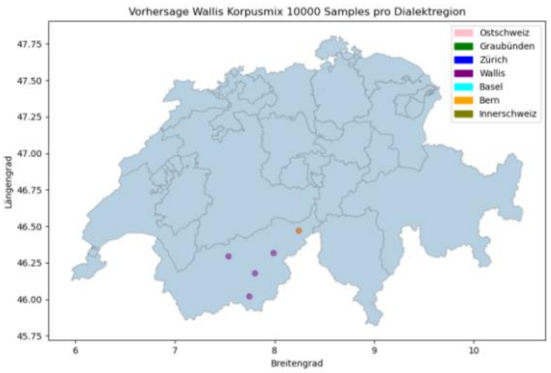


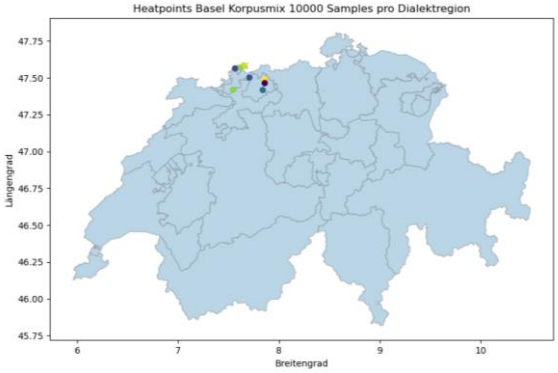


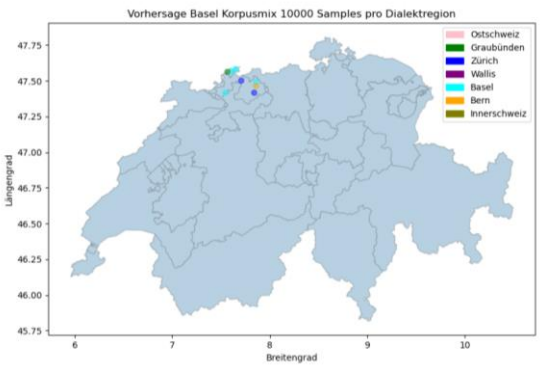

Heatpoints Zürich Korpusmix 10000 Samples pro Dialektregion
Längengrad
Breitengrad

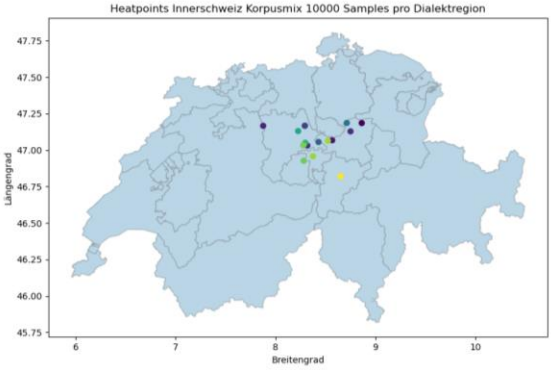

Vorhersage Zürich Korpusmix 10000 Samples pro Dialektregion
Ostschweiz
Graubünden
Zürich
Wallis
Basel
Bern
Innerschweiz
Längengrad
Breitengrad

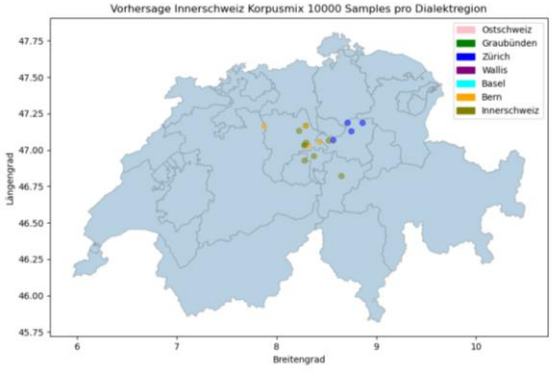

Heatpoints Wallis Korpusmix 10000 Samples pro Dialektregion
Längengrad
Breitengrad


Vorhersage Wallis Korpusmix 10000 Samples pro Dialektregion
Ostschweiz
Graubünden
Zürich
Wallis
Basel
Bern
Innerschweiz
Längengrad
Breitengrad

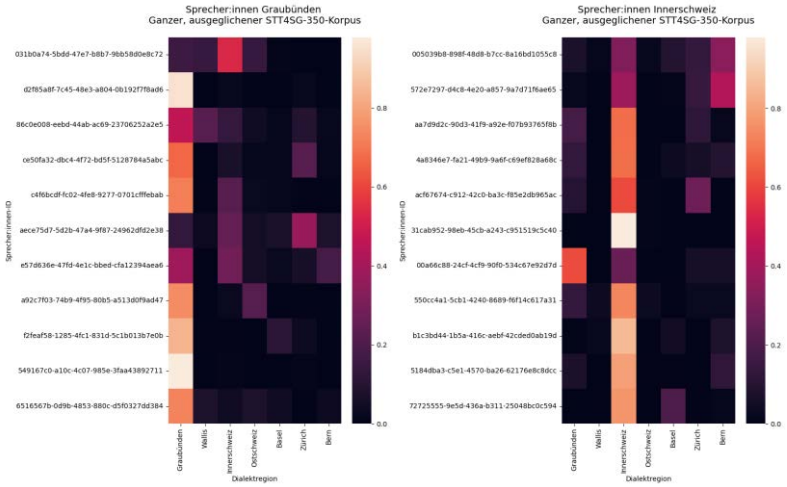

Heatpoints Basel Korpusmix 10000 Samples pro Dialektregion
Längengrad
Breitengrad

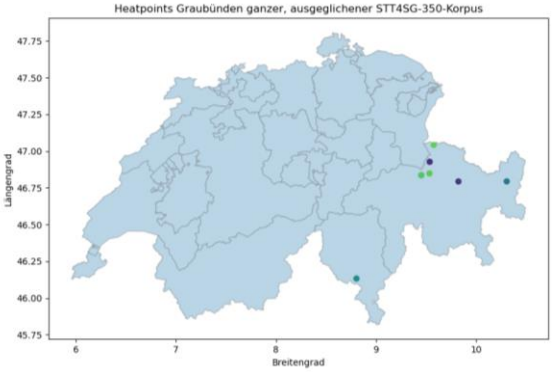

Vorhersage Basel Korpusmix 10000 Samples pro Dialektregion
Ostschweiz
Graubünden
Zürich
Wallis
Basel
Bern
Innerschweiz
Längengrad
Breitengrad

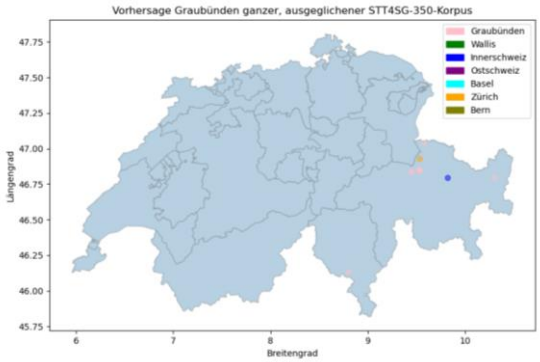

Heatpoints Innerschweiz Korpusmix 10000 Samples pro Dialektregion
Längengrad
Breitengrad

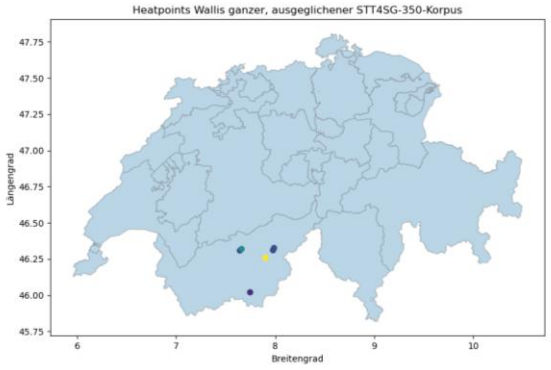

Vorhersage Innerschweiz Korpusmix 10000 Samples pro Dialektregion
Ostschweiz
Graubünden
Zürich
Wallis
Basel
Bern
Innerschweiz
Längengrad
Breitengrad

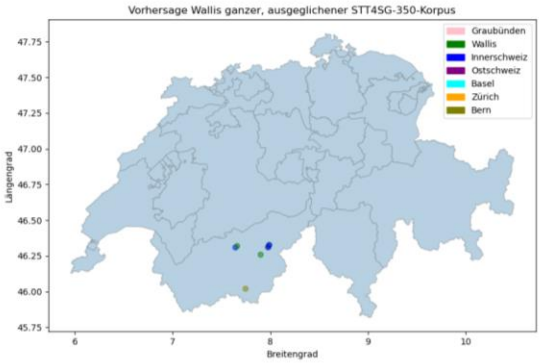
6.2.2 Experiment 2

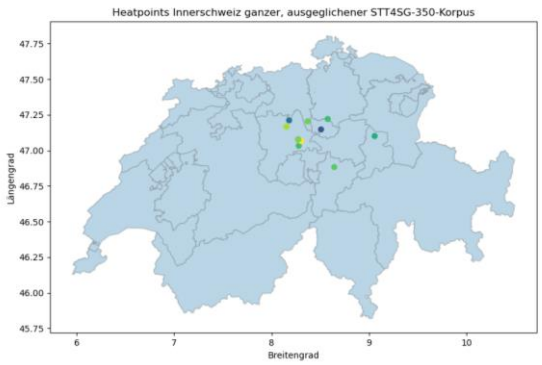


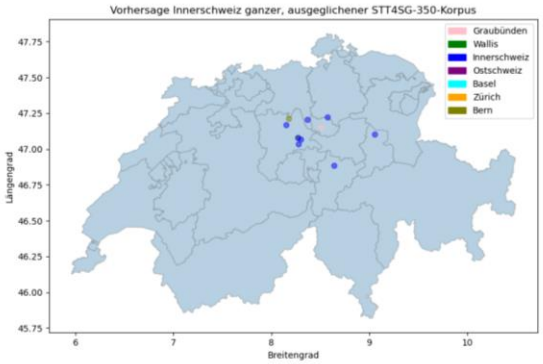


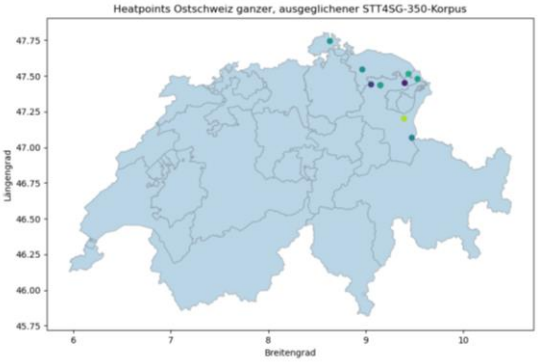


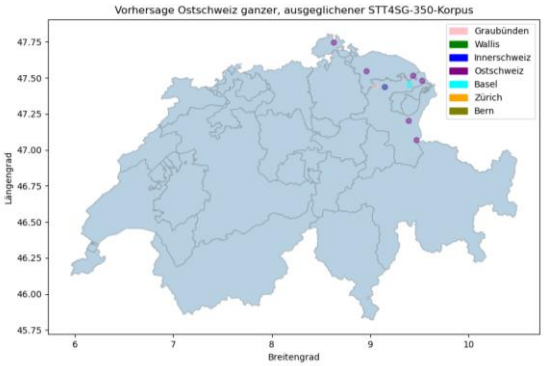


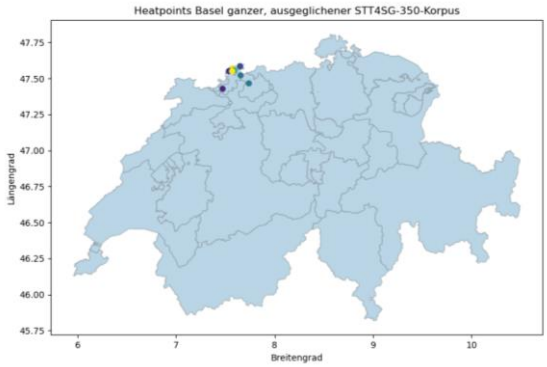


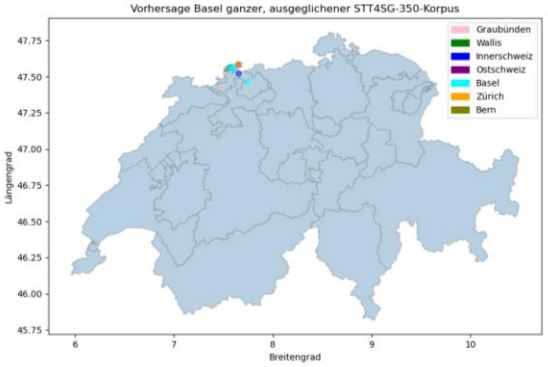

Heatpoints Innerschweiz ganzer, ausgeglichener STT4SG-350-Korpus
Längengrad
Breitengrad

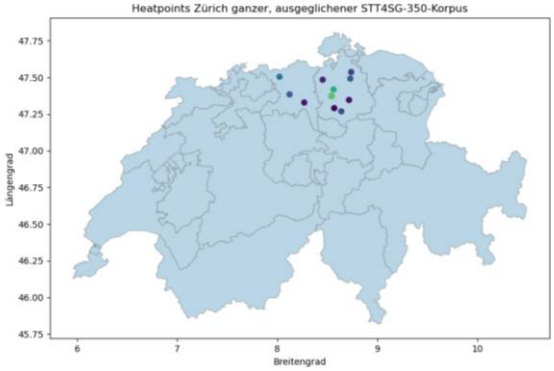

Vorhersage Innerschweiz ganzer, ausgeglichener STT4SG-350-Korpus
Graubünden
Wallis
Innerschweiz
Ostschweiz
Basel
Zürich
Bern
Längengrad
Breitengrad

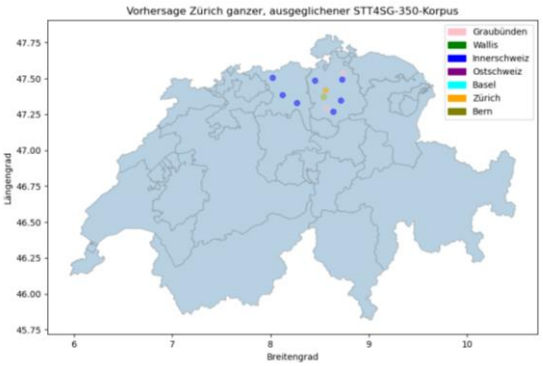

Heatpoints Ostschweiz ganzer, ausgeglichener STT4SG-350-Korpus
Längengrad
Breitengrad

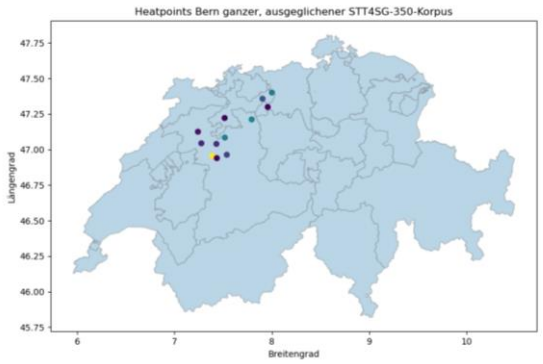

Vorhersage Ostschweiz ganzer, ausgeglichener STT4SG-350-Korpus
Graubünden
Wallis
Innerschweiz
Ostschweiz
Basel
Zürich
Bern
Längengrad
Breitengrad

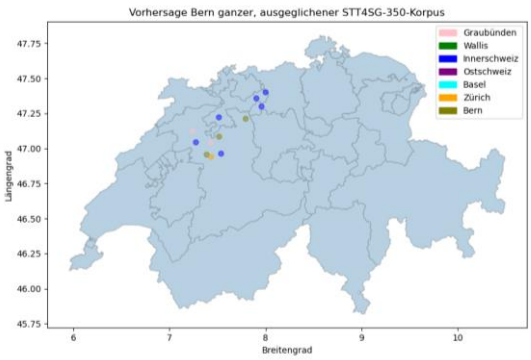

Heatpoints Basel ganzer, ausgeglichener STT4SG-350-Korpus
Längengrad
Breitengrad


Vorhersage Basel ganzer, ausgeglichener STT4SG-350-Korpus
Graubünden
Wallis
Innerschweiz
Ostschweiz
Basel
Zürich
Bern
Längengrad
Breitengrad

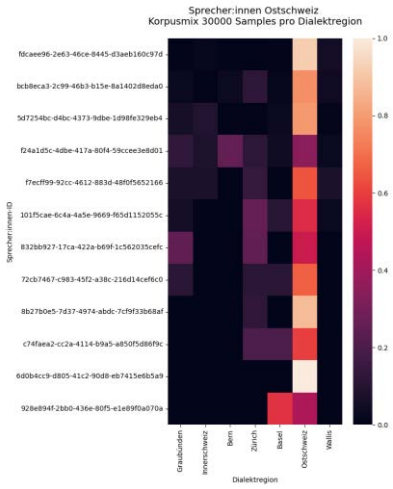

Heatpoints Zürich ganzer, ausgeglichener STT4SG-350-Korpus
Längengrad
Breitengrad

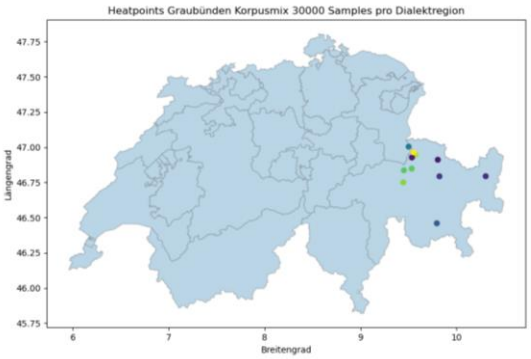

Vorhersage Zürich ganzer, ausgeglichener STT4SG-350-Korpus
Graubünden
Wallis
Innerschweiz
Ostschweiz
Basel
Zürich
Bern
Längengrad
Breitengrad

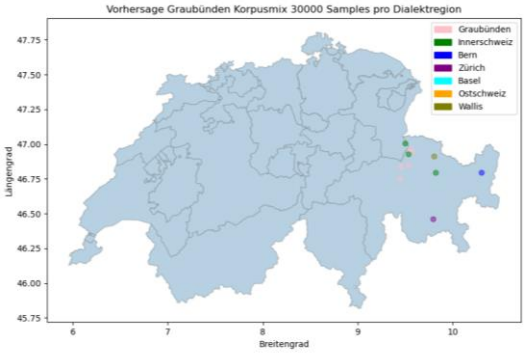


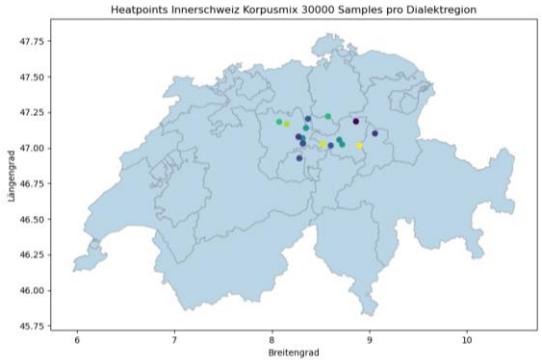


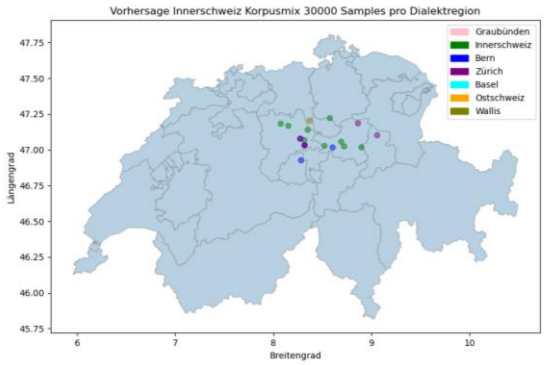
6.2.3 Experiment 5

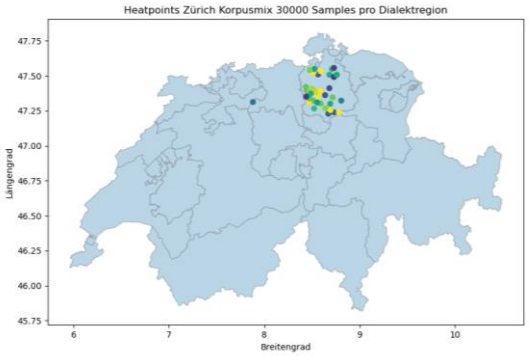


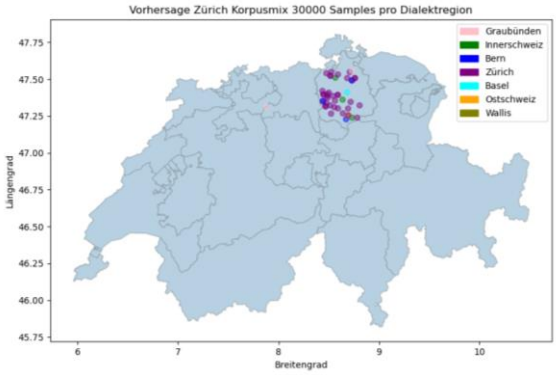


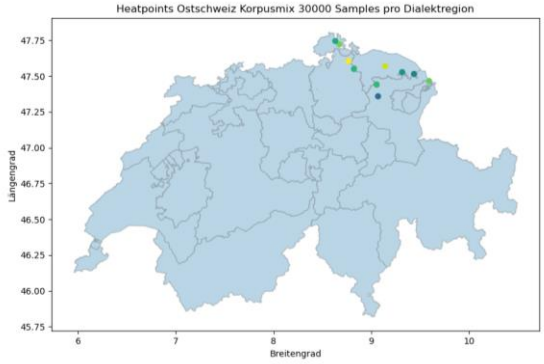


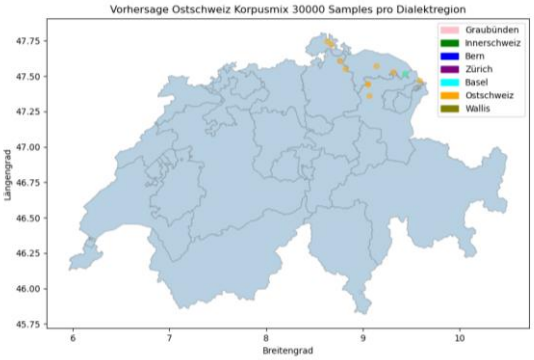

Heatpoints Innerschweiz Korpusmix 30000 Samples pro Dialektregion
Längengrad
Breitengrad

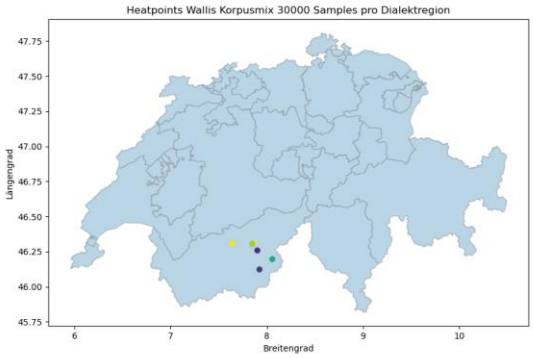

Vorhersage Innerschweiz Korpusmix 30000 Samples pro Dialektregion
Graubünden
Innerschweiz
Bern
Zürich
Basel
Ostschweiz
Wallis
Längengrad
Breitengrad

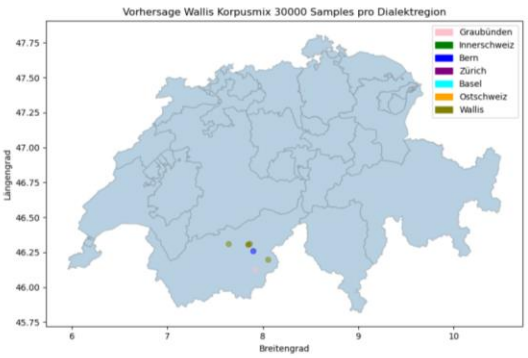

Heatpoints Zürich Korpusmix 30000 Samples pro Dialektregion
Längengrad
Breitengrad


Vorhersage Zürich Korpusmix 30000 Samples pro Dialektregion
Graubünden
Innerschweiz
Bern
Zürich
Basel
Ostschweiz
Wallis
Längengrad
Breitengrad


Heatpoints Ostschweiz Korpusmix 30000 Samples pro Dialektregion
Längengrad
Breitengrad


Vorhersage Ostschweiz Korpusmix 30000 Samples pro Dialektregion
Graubünden
Innerschweiz
Bern
Zürich
Basel
Ostschweiz
Wallis
Längengrad
Breitengrad


Heatpoints Wallis Korpusmix 30000 Samples pro Dialektregion
Längengrad
Breitengrad
47.75
47.50
47.25
47.00
46.75
46.50
46.25
46.00
45.75
6
7
8
9
10


Vorhersage Wallis Korpusmix 30000 Samples pro Dialektregion
Graubünden
Innerschweiz
Bern
Zürich
Basel
Ostschweiz
Wallis
Längengrad
Breitengrad
47.75
47.50
47.25
47.00
46.75
46.50
46.25
46.00
45.75
6
7
8
9
10

6.3 Übersicht über die Trainingsdurchläufe

Pretrained Model	Run Name	Datum	Dauer	Dataset Train	Dataset Valid	Dataset Test	Batch Size	Gradient Accumulation Factor	Epochs	Label
wav2vec	Nimbus 2003	05.11.2023	4d	Mix 10000 Samples pro Region	Mix 10000 Samples pro Region	Mix 10000 Samples pro Region	2	4	20	Dialect Region
whisper	Fluffy	13.11.2023	08:55:34	Mix 10000 Samples pro Region	Mix 10000 Samples pro Region	Mix 10000 Samples pro Region	8	4	5	Dialect Region
whisper	Fluffy20	17.11.2023	1d 6:25	Mix 10000 Samples pro Region	Mix 10000 Samples pro Region	Mix 10000 Samples pro Region	8	4	20	Dialect Region
whisper	Seidenschnabel	19.11.2023	4d 00:36:21	Train balanced von Korpus	Valid von Korpus	Test von Korpus	8	4	20	Dialect Region
whisper	Umbridge	24.11.2023	07:26:42	Unausgeglichene Splits	Unausgeglichene Splits	Unausgeglichene Splits	8	4	20	Dialect Region Englisch
whisper	Snape	26.11.2023	07:26:42	Ausgeglichene Splits	Ausgeglichene Splits	Ausgeglichene Splits	8	4	20	Dialect Region Englisch
whisper	Krummbein	28.11.2023	1d 01:23:51	Ausgeglichen Anzahl Samples pro Sprecher:in	Ausgeglichen Anzahl Samples pro Sprecher:in	Ausgeglichen Anzahl Samples pro Sprecher:in	8	4	20	Dialect Region
whisper	Fawkes	11.12.2023	4d 02:12:39	Mix 30000 Samples pro Region	Mix 30000 Samples pro Region	Mix 30000 Samples pro Region	8	4	20	Dialect Region